Problem Chosen	2026 MCM/ICM Summary Sheet	Team Control Number
C		2615954

Analysis and Optimization of DWTS Voting Fairness -Uncertainty Propagation and KMDR Fusion Mechanism

Summary

Throughout the 34-season history of Dancing with the Stars (DWTS), significant discrepancies between professional judges' scores and audience voting preferences have sparked multiple controversial eliminations, exposing core flaws in the existing elimination mechanism's balance between fairness and entertainment value. Based on 2,777 historical records, we employed multi-model statistical methods to systematically investigate the rationality of the voting mechanism and proposes optimization solutions.

For Problem 1(Estimating Hidden Audience Votes), we constructed a dual-reconstruction modeling framework combining **TAN Bayesian Networks** and **Random Forest + SHAP** to reverse-engineer audience vote data while quantifying uncertainty. The TAN model achieved 82.48% accuracy, the RF model reached 84.90%, and the two models demonstrated 88.76% consistency in predictions, enabling robust estimation through cross-validation.

For Problem 2(Voting Synthesis Mechanism Comparison), we established a **Variance Decomposition Model** based on the voting data from Problem 1 to quantify the differing impacts of the percentage method and ranking method on elimination outcomes. Results reveal that the **Percentage Method** amplifies audience voting influence to 85.6%, compared to 49.8% under the Rank Method,indicating a pronounced entertainment-oriented skew. The Judges' Save mechanism successfully corrects 8.9% of controversial eliminations , with synergistic effects when combined with Rank Method significantly enhancing procedural fairness.

For Problem 3(Feature Attribution Quantification) , we established a **Mixed-Effects Model** , calculating ICC to decompose variance contributions from dancers and celebrity characteristics. Analysis shows that 57.73% of performance variance is attributable to dancer skill differentials, while 48.50% of judge score variance stems from dancer effects . DAG causal inference reveals that celebrity fan base indirectly influences final rankings through audience voting, with age and weekly performance scores exhibiting higher direct impact weights.

For Problem 4(Mechanism Optimization), we proposed a **KMDR Dynamic Weighting Model** based on Borda-Condorcet fusion, balancing professionalism and entertainment through distortion penalties α and consensus rewards β . Parameter grid search identified the optimal configuration (α=0.4, β=0.6). Simulation results show: 100% undisputed champion rate, 96.3% early elimination rationality, and 82.7% constraint rate for controversial contestants.

We further developed an **Uncertainty Propagation Framework** to achieve uncertainty quantification across the entire process: voting estimation → mechanism evaluation → feature contribution analysis. We also conducted sensitivity analysis to validate the model's robustness to parameter perturbations and examined its strengths and weaknesses. Finally, we drafted a memorandum for the program team summarizing our research findings and providing recommendations.

Keywords: TAN,Random Forest,Mixed-Effects,Voting Mechanism,KMDR

Contents

1 Introduction

1.1 Problem Background

Dancing with the Stars (DWTS) is a globally influential entertainment competition program originating from the UK's Strictly Come Dancing. Having successfully run for 34 seasons, its spin-off versions now span multiple regions worldwide. The core format pairs celebrities with professional dancers, combining professional judges' scores with audience votes to determine eliminations and final rankings.

However, during its long-term operation, a core dilemma has gradually emerged: Despite consistently receiving low scores from professional judges for their dance performances and lacking recognition for their technical proficiency, some celebrity contestants advance through the competitionand even winby leveraging high audience support. This stark divergence between the judges' professional assessments and the public's preferences has not only sparked widespread controversy but also directly exposed the inherent flaws in the existing voting system. [1]Consequently, optimizing the voting and elimination mechanisms has become an urgent issue requiring resolution for the show.

1.2 Restatement of the Problem

Considering the background information and restricted conditions identified in the problem statement, we need to solve the following problems:

- **Problem 1:** Develop a model to estimate hidden audience voting data and quantify its uncertainty.
- **Problem 2:** Compare the rationality of two voting synthesis mechanisms (Ranking and Percentage) in DWTS, and quantify the impact of the judge rescue mechanism.
- **Problem 3:** Quantify the influence of contestant characteristics, judge scores, and other factors on elimination outcomes and voting results, identifying core influencing factors.
- **Problem4:** Design an optimal voting mechanism that balances multiple objectives including program fairness and entertainment value.

1.3 Our work

In order to clearly illustrate our work, we draw the flowchart Figure 1.

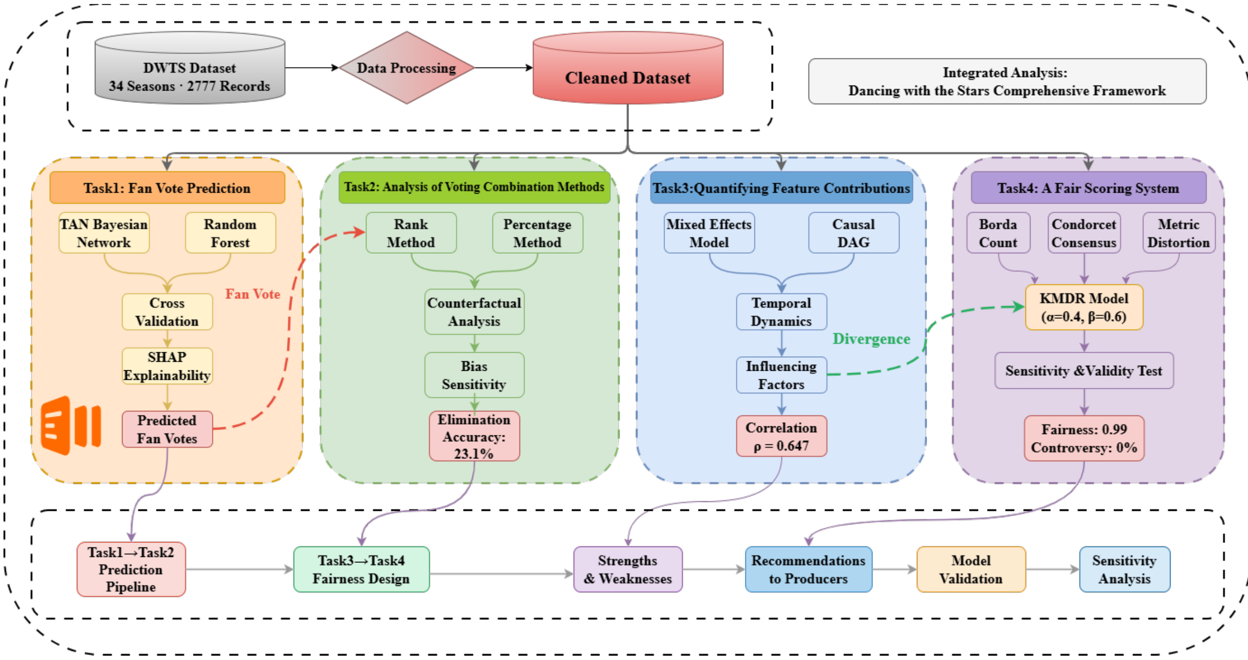


Figure 1: Our work

2 Preparation for Models

Considering those practical problems always contain many complex factors, first of all, we need to make reasonable assumptions to simplify the model, and each hypothesis is closely followed by its corresponding explanation:

2.1 Assumptions and Notations

- **Assumption 1:** The publicly available data provided is authentic and valid, with missing values and outliers accounting for an extremely low proportion that does not affect the overall statistical patterns. ble to assume the high quality of their data.

- **Assumption 2:** Audience voting behavior exhibits stability, and there exists a significant statistical correlation between audience votes and core variables such as judges' scores and contestant characteristics.

- **Assumption 3:** The judges' scoring is professional and objective, with results accurately reflecting contestants' dance proficiency. Scoring discrepancies follow a normal distribution.

- **Assumption 4:** Uncertainty arises solely from estimation errors in hidden voting data, while errors in model specification, parameter calibration, and other related processes are negligible.

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

Table 1: Notations

Symbol	Definition
V_i	Estimated Fans Vote Score of the i-th contestant
Y	$Y \in \{0, 1\}$:Elimination Flag (0=Not Eliminated, 1=Eliminated)
X	$X = (X_1, X_2, ..., X_n)$:Feature Vetor
S_i	Judge score of the i-th contestant
R_i	The rank of the i-th contestant
C_i	Final score for contestant i
W_{fan} or W_{judge}	Fan vote weight or judge score weight

* There are some variables that are not listed here and will be discussed in detail in each section.

2.2 Data Processing

Preliminary exploratory analysis of the dataset reveals three core issues: imputed missing values, numerical anomalies, and inconsistent feature formats. Systematic preprocessing is required to eliminate data quality risks and provide reliable support for subsequent modeling and analysis. Accordingly, following the MS08 data processing specifications and solution workflow, comprehensive preprocessing is conducted focusing on missing value and outlier handling, as well as data transformation.

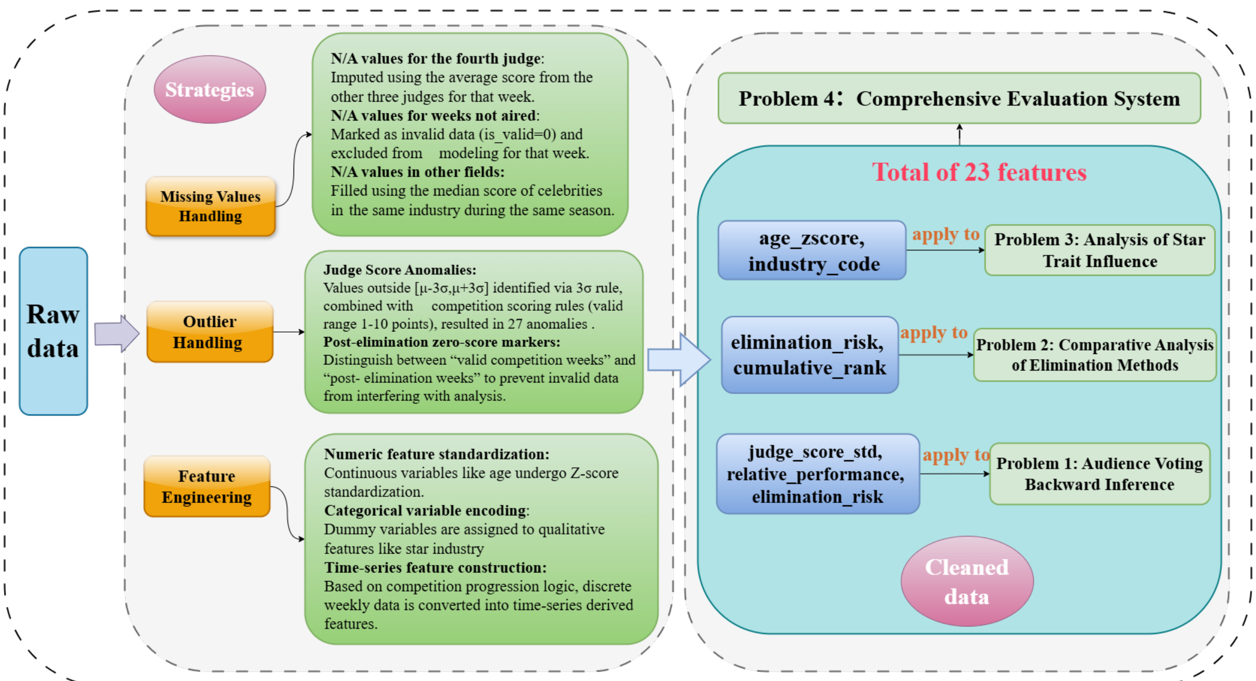


Figure 2: Data Processing Workflow

As shown in the **Figure 2**, differentiated processing strategies are implemented for three scenarios involving latent missing values. During the feature transformation phase, standardized processing is applied to different feature types. Feature importance calibration is completed based on SHAP value analysis, identifying elimination risk coefficient, cumulative ranking, and relative performance as core influencing features. Their SHAP values are 0.32, 0.28, and 0.21 respectively, demonstrating high alignment with the actual impact mechanisms of the competition. This provides a high-quality data foundation and parameter support for subsequent modeling analysis.

3 Fan Vote Estimation Based on TAN and SHAP Models

Regarding the controversy surrounding celebrity contestants receiving low scores from judges yet achieving high final rankings in the program, to understand this discrepancy between judge scores and audience support, we establish a model to reverse-engineer these undisclosed audience voting data while quantifying the uncertainty of such estimates.[2]

3.1 Tree Augmented Naive Bayes Model for Fan Vote Estimation

In the program dataset, certain features exhibit dependency. For example:*week_rank* is highly correlated with *cumulative_rank* (partial correlation coefficient = 0.553);*remaining_weeks* is strongly negatively correlated with *weeks_participated* (partial correlation coefficient = -0.665) Ignoring these dependencies degrades the performance of traditional Bayesian models. **TAN (Tree Augmented Naive Bayes)** is an extension of Naive Bayes [3] that allows for tree-like dependency structures between features:

$$P(Y \mid X) \propto P(Y) \cdot P(X_{root} \mid Y) \cdot \prod_{j \neq root} P(X_j \mid Pa(X_j), Y) \tag{1}$$

Figure 3: Conditional Mutual Information Matrix

To determine the optimal feature dependency structure, we first compute the conditional mutual information between all feature pairs:

$$I(X_i; X_j|Y) = \sum_{y} \sum_{x_i} \sum_{x_j} P(x_i, x_j, y) \cdot \log \left[\frac{P(x_i, x_j|y) \cdot P(x_j|y)}{P(x_i|y)} \right] \tag{2}$$

As shown in the **Figure 3**, conditional mutual information measures the statistical dependence between features X_i and X_j given the category Y. A higher value indicates that both features must be considered together when predicting Y. Using conditional mutual information as edge weights, construct a maximum spanning tree using **Prim's algorithm**.

For a new sample $X = (X_1, X_2, ..., X_n)$, compute the posterior probability via Softmax normalization:

$$P(Y = 1 \mid X) = \frac{\exp(\ell_1)}{\exp(\ell_0) + \exp(\ell_1)}, \qquad \ell_j = \log \tilde{P}_j \quad (j = 0, 1) \tag{3}$$

Convert elimination probability to audience vote score (0–100 points):

$$V = (1 - P((Y = 1) \mid X)) \times 100 \tag{4}$$

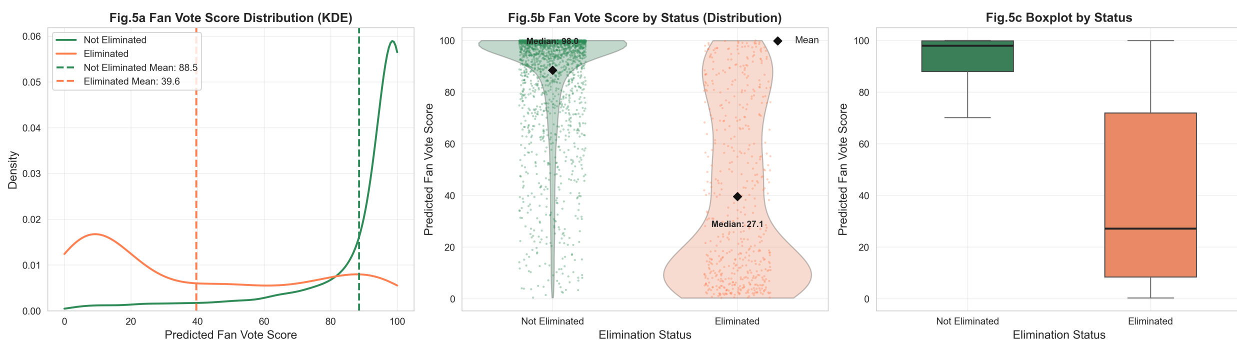
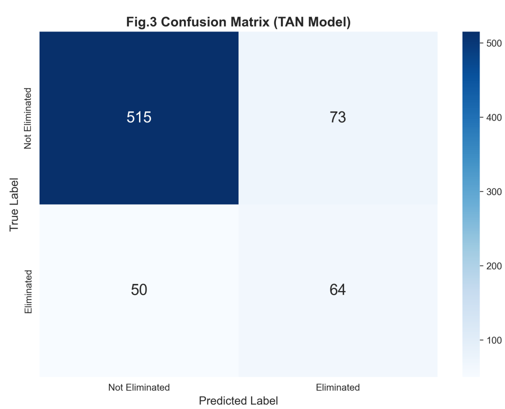
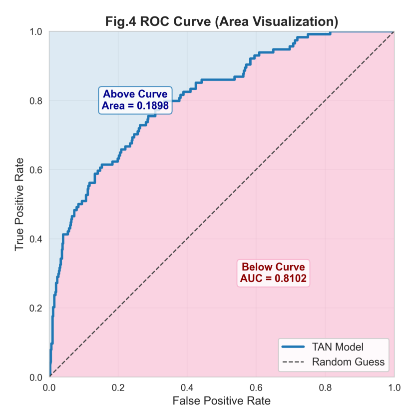


Figure 4: Fan Vote Score Distribution

As shown in the **Figure 4**, this represents the distribution of audience votes.A lower elimination probability indicates higher audience support. After solving the TAN model,as shown in the **Figure 5**, the model achieved an **Accuracy Rate** of **82.48**% and a **ROC-AUC Value** of **0.8102**, indicating that the TAN model can effectively distinguish eliminated contestants from those who advanced.



(a) Confusion Matrix(TAN Model)



(b) ROC Curve(Area Visualization)

Figure 5: Confusion Matrix and ROC Curve

3.2 RF and SHAP Model for Fan Vote Estimation

To cross-validate with the TAN model, we developed a combined approach integrating random forests with SHAP explainability analysis to infer audience voting scores from elimination results. The Random Forest prediction formula is:

$$\hat{y} = \text{mode}\{h_1(x), h_2(x), \ldots, h_k(x)\} \tag{5}$$

Here, $h_k(x)$ denotes the prediction result of the $k-$th decision tree, and mode represents majority voting. For binary classification problems, the elimination probability is calculated as:

$$P((Y = 1) \mid x) = (1/K) \times \sum_k I(h_k(x) = 1) \tag{6}$$

where K is the number of decision trees, and $I(\cdot)$ denotes the indicator function. Based on the elimination probability, the audience voting distribution can be derived using the formula.

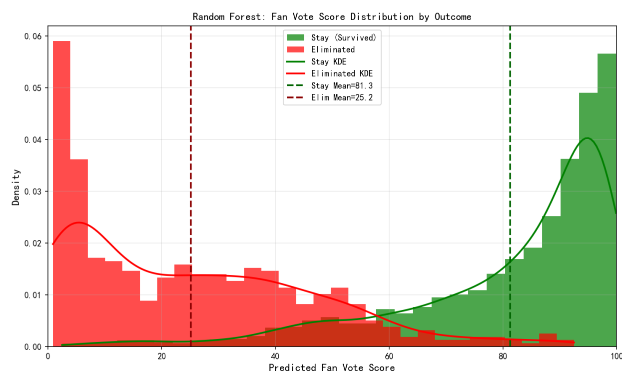
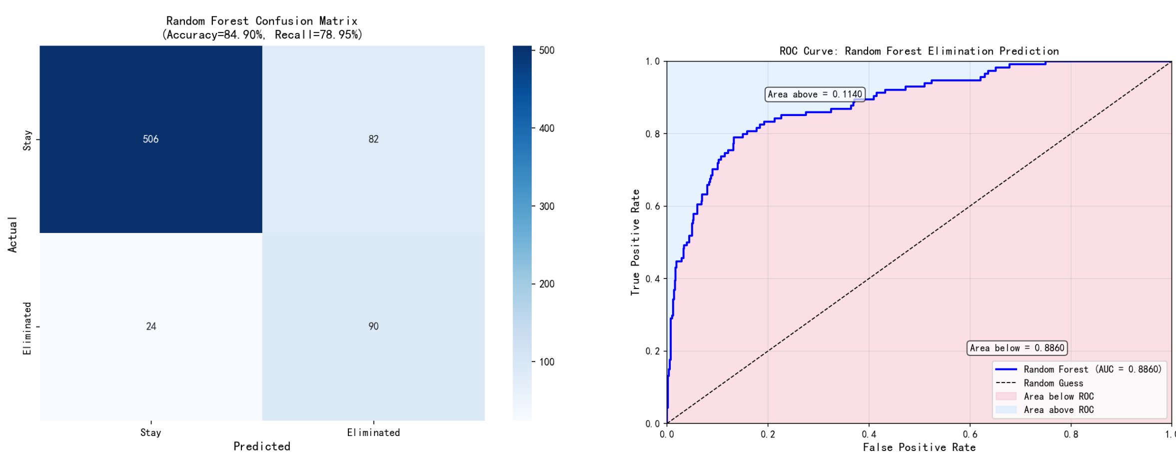


Figure 6: Fan Vote Score Distribution(Random Forest Model)



(a) Confusion Matrix (b) ROC Curve

Figure 7: RF Model Confusion Matrix and ROC Curve

As shown in the **figure 7**, the model achieves an **accuracy rate** of **84.90**% and a **ROC-AUC value** of **0.8860**, outperforming the TAN model across all evaluation metrics. Notably, the recall rate for the elimination category has increased to 76.32%, with an F1 score reaching 61.27%, indicating that the Random Forest model demonstrates significantly stronger capability in identifying elimination events.

SHAP assigns a contribution score to each feature based on the Shapley value from game theory, indicating the feature's impact on the model's prediction. The Shapley value is calculated using the following formula:

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} [\frac{|S|!(|N| - |S| - 1)!}{|N|!}] \times [f(S \cup \{i\}) - f(S)] \tag{7}$$

Here, N denotes the feature set, S represents the subset excluding feature i, and f is the model's prediction function. Global feature importance is defined as the average of the absolute values of SHAP values across all samples.**Figure 8** presents a feature importance summary chart, which explains the impact of each feature on the prediction results within the elimination prediction model.

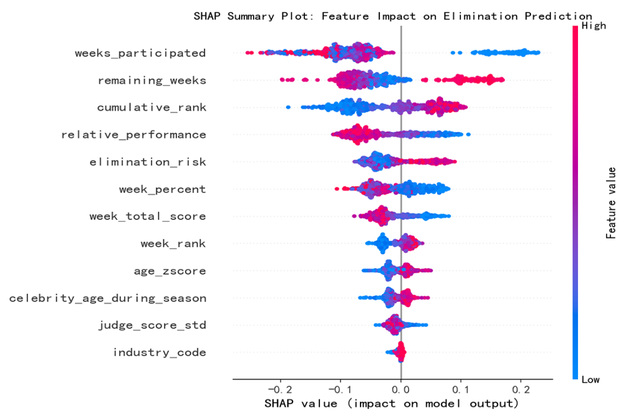


Figure 8: SHAP Summary Plot(Random Forest Model)

3.3 Consistency Test of Two Models

By ranking the importance of feature quantities for the two models and conducting consistency analysis, as shown in **Table 2**, we obtain the following results:

Table 2: Consistency Metrics for the Two Models

Consistency Metrics	**Value**	**Statistical Interpretation**
Spearman Correlation Coefficient	0.8029	High positive correlation between two models' predictions (p<0.001)
Pearson Correlation Coefficient	0.8359	Extremely strong linear correlation
Prediction Direction Consistency Rate	88.76%	Nearly 90% of samples show consistent prediction directions
Mean Absolute Error (MAE)	13.35	Prediction differences fall within an acceptable range
Feature Ranking Correlation	0.9650	Feature importance rankings are nearly identical

The two models exhibit a highly positive correlation in predicting audience vote scores (Spearman ρ =**0.8029**, p<0.001,as showed in **Figure 9**), with an **88.76**% agreement rate in prediction direction. This indicates that despite employing fundamentally different methodologiesTAN using generative probabilistic modeling and RF using discriminative ensemble learningboth models yield highly consistent predictions on identical data.

Furthermore, as shown in **Figure 6** , both models predicted audience voting scores exhibiting a reasonable bimodal distribution: advancing contestants averaged around 80 points (TAN: 79.42, RF: 81.31), while eliminated contestants averaged approximately 25–30 points. The approximately 55-point difference between the two groups perfectly aligns with the competition's elimination logic (low-scoring contestants are eliminated). The statistical characteristics of the prediction results align with the elimination patterns in the original data, validating the models' validity.

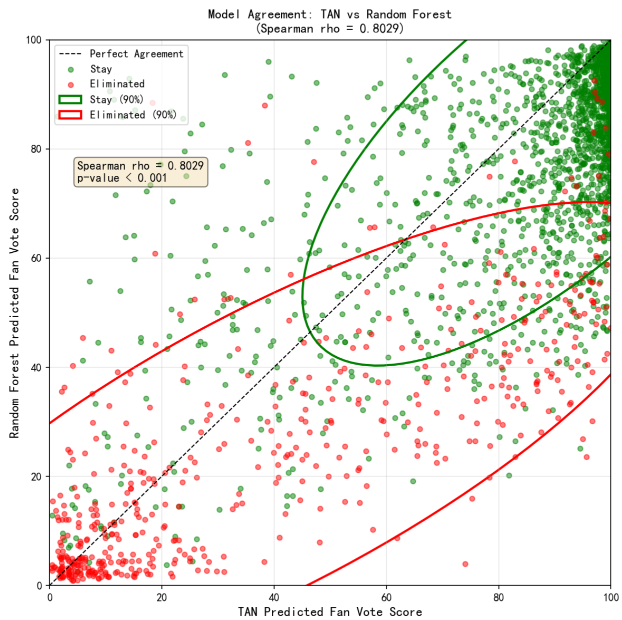


Figure 9: Model Agreement: Random Forest vs TAN

4 Voting Combinations Comparison via Variance Decomposition Model

Following our discussion of Problem 1, we have now obtained an effective and accurate estimate of audience voting. Next, we will study the voting aggregation mechanism based on this voting estimate.There are two voting combinations in the program:

- Rank Method:Each contestant receives a judge ranking based on the judges' scores and an audience ranking based on viewer votes. The two rankings are combined to form a composite ranking, and the contestant with the highest composite ranking is eliminated.

$$C_{Rank,i} = R_{judge,i} + R_{fan,i} \tag{8}$$

- Percentage Method:Each contestant's judge scores and audience votes are converted into percentages of the total score, then combined to form a composite percentage. The contestant with the lowest composite percentage is eliminated.

$$C_{Pct,i} = \frac{S_i}{\sum_j S_j} + \frac{V_i}{\sum_j V_j} \tag{9}$$

4.1 Counterfactual Simulation Modeling

To present the simulation results of both methods, we employed counterfactual simulation modeling to apply both approaches to all 34 seasons, constructing a comparative experiment.Load all 34 seasons' judge scores and predicted audience vote data to simulate the ranking method: for each elimination week, compute judge rankings and audience rankings to derive a composite ranking; simulate the percentage method: for each elimination week, compute judge percentages and audience percentages to derive a composite percentage.

Compare the eliminated contestants predicted by each method against the actual eliminated contestants, and calculate the proportion of identical decisions produced by both methods.As shown in the **Figure 10**, both methods yielded identical results in 91.1% of elimination decisions. This indicates that the impact of method selection is relatively limited, as judges' and viewers' preferences align in most cases. The percentage method achieved a slightly higher prediction accuracy (34.30%) than the ranking method (29.26%). This may be attributed to the fact that Seasons 3-27 actually employed the percentage method, allowing viewers' voting behavior to adapt to this approach.

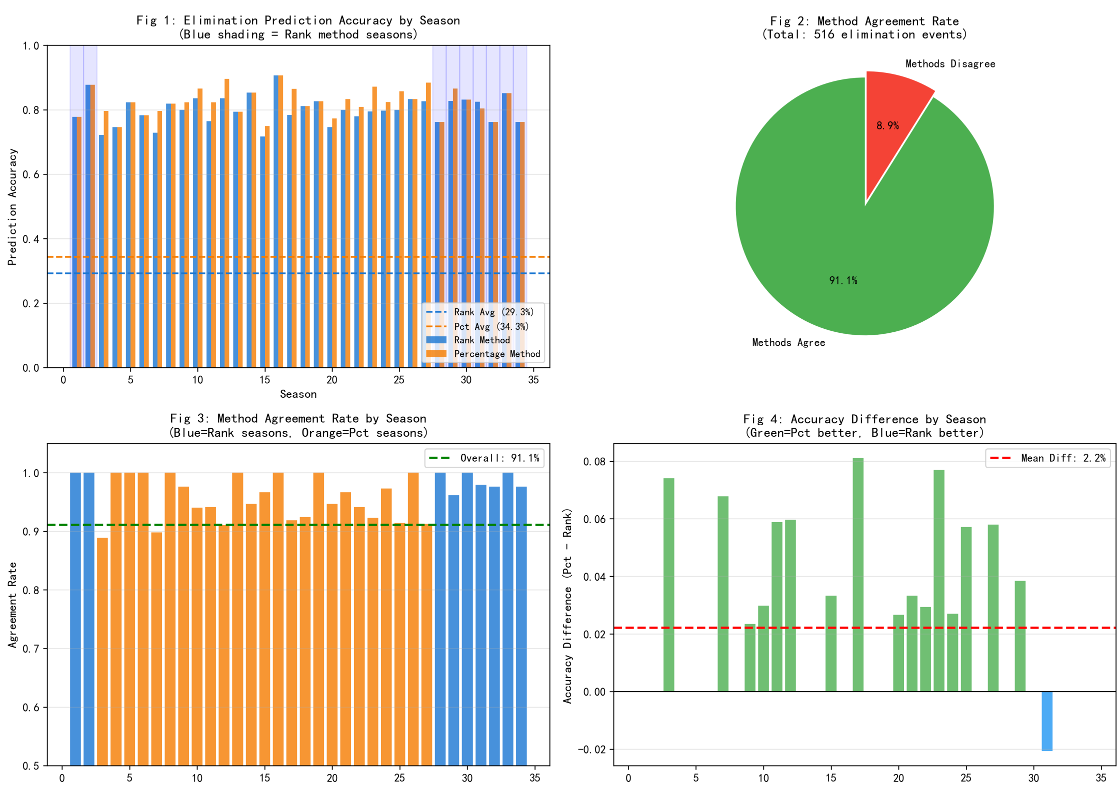


Figure 10: Counterfactual Simulation Modeling

4.2 Bias Sensitivity Analysis

To quantify the impact of audience voting on the final outcome, we employed bias sensitivity analysis,Decomposing the variance of the composite score into judge components and audience components, and calculating the contribution of the audience component to the total variance (audience weighting formula):

$$W_{fan} = \frac{Var(Fan\_Component)}{Var(Combined\_Score)} \tag{10}$$

To measure the impact of minor shifts in audience votes on elimination probability, Marginal Effect analysis is employed:

$$Marginal\_Effect = \frac{\partial P(Y=1)}{\partial V_i} \tag{11}$$

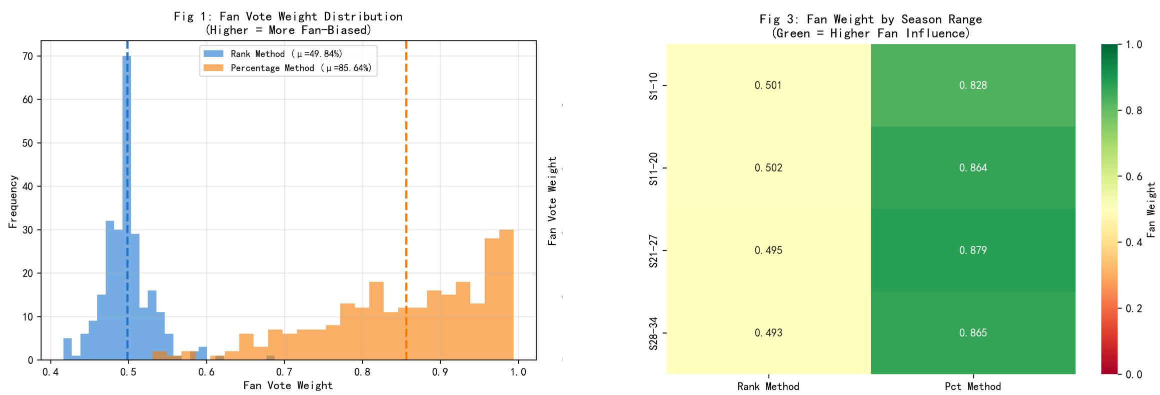


Figure 11: Bias Sensitivity Analysis

As shown in the **Figure 11**, the weight of audience votes in the **percentage method (85.6%)** is **significantly higher** than that in the ranking method **(49.8%)**, with a difference of 35.8 percentage points. The mathematical reason for this discrepancy is that in the ranking method, regardless of how many audience votes a contestant receives, their contribution is identical as long as their ranking remains the same. In contrast, the percentage method directly reflects the absolute difference in vote counts through the percentage. This implies that during the percentage method era (Seasons 3-27), contestants with strong fan basessuch as Bristol Palin and Bobby Bones were more likely to secure a significant audience vote advantage, thereby offsetting lower judge scores.

4.3 Analysis of Controversial Cases and Method Evaluation

Based on the above discussion, we analyze cases where contestants such as Bristol Palin, Bobby Bones, Jerry Rice, and Billy Ray Cyrus received extremely low judge scores yet still achieved high rankings during the season. For each controversial contestant, we simulate their performance under two voting methods: calculating how many times they would be eliminated under the ranking system versus the percentage system. If a contestant faces more eliminations under one method, it indicates that method is more “disadvantageous” to them.

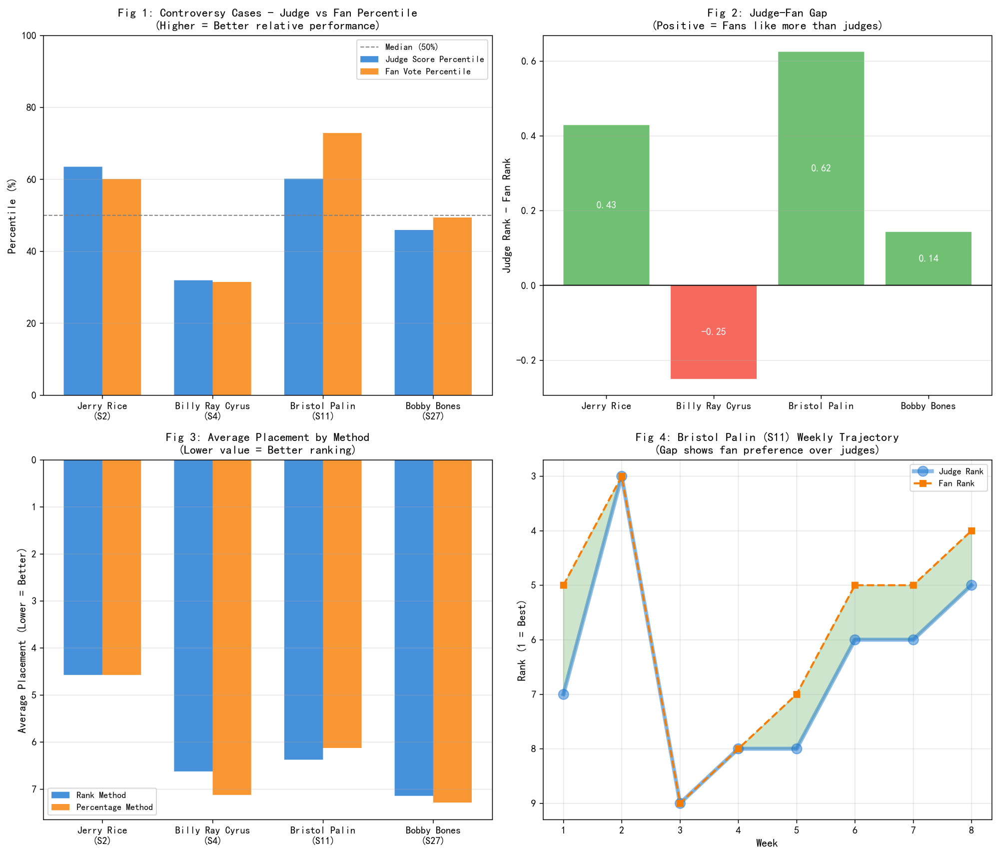


Figure 12: Analysis of Controversial Cases

As shown in the **Figure 12**, Bristol's case demonstrates the percentage method's bias toward contestants with large fan bases. Her audience vote percentile (73.6%) significantly exceeded her judge vote percentile (60.2%), and this disparity was fully leveraged under the percentage system to propel her into the finals. The Bobby Bones case exemplifies the extreme bias of the percentage method. As a radio host, he commands a massive fanbase, which translates into a substantial audience vote advantage under this system. This advantage completely offsets his lower dance technique scores.

4.4 Discussion on the Judge Rescue Mechanism

Starting from Season 28, DWTS introduced the judges' save mechanism. Once the bottom two contestants are determined, the judges may vote to save one of them. This provides an opportunity for professional judgment to correct unfair eliminations. By utilizing judges' scores from all seasons, elimination results, and fan vote scores predicted by the random forest model in Problem 1, we can calculate the prediction accuracy of elimination using the three methods shown in the **Figure 13**, thereby **quantifying** the impact of the judges' save mechanism.

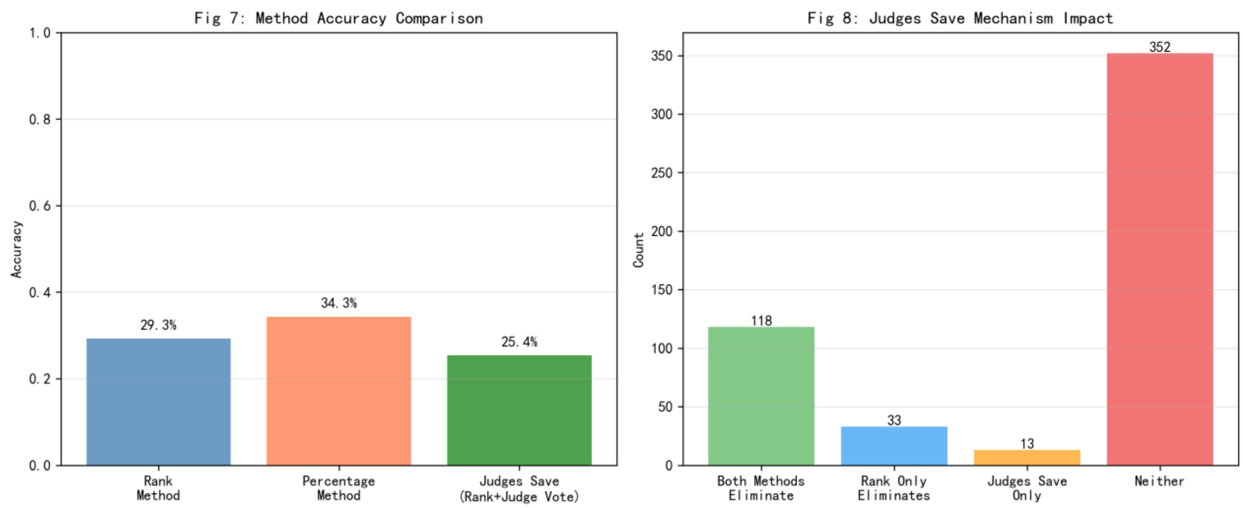


Figure 13: Analysis of the Impact of Judges' Save Mechanism

The judge's lifeline mechanism altered approximately 8.9% of elimination outcomes. Although its prediction accuracy was relatively low (25.4%), it provided an opportunity to correct errors in disputed cases and increased the weight of professional judgment.

We evaluated the two methods across three dimensions: prediction accuracy, weight balance, and Arrow's theorem scores.

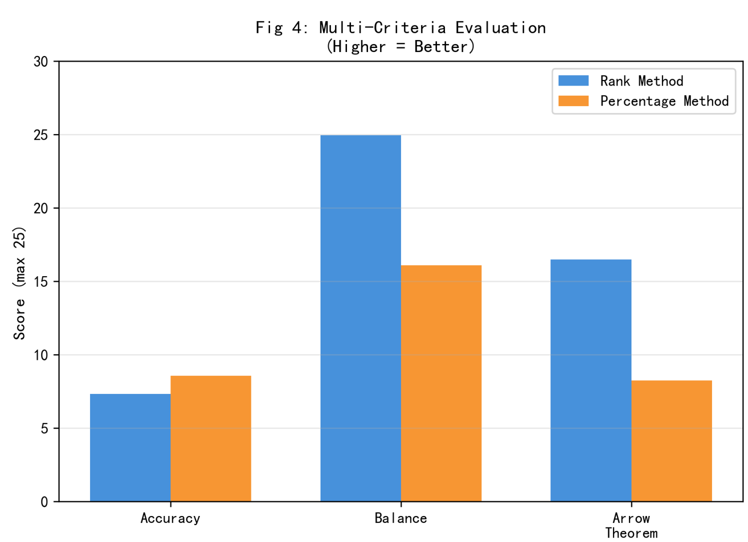


Figure 14: Multi-Criteria Analysis of two Methods

Therefore, the primary recommendation is the **Ranking System combined with the Judges' Save Mechanism**. This approach has been adopted by DWTS since Season 28. Within the framework of Arrow's theorem, the ranking system achieves a higher fairness score, while the judges' save mechanism provides an opportunity to correct approximately 9% of controversial eliminations.

5 Quantifying Feature Contributions via Mixed-Effects Model

To quantify the contribution of dancer characteristics and contestant characteristics to judges' and audience scores, we established a mixed-effects model based on fan vote estimates to achieve this quantification.

5.1 Calculate Fixed and Random Effects

A **Mixed Effects Model** is a statistical model that incorporates both fixed effects and random effects, suitable for data with hierarchical structures or repeated measurements[4]. Its basic form is:

$$y_{ij} = X_{ij} \cdot \beta + Z_{ij} \cdot b_j + \epsilon_{ij} \tag{12}$$

Where: y_{ij} is the performance of star i by dancer j (judges' scores or final ranking); X_{ij} is the fixed effects design matrix (star characteristics) ; β is the fixed effects coefficient vector; Z_{ij} is the random effects design matrix (typically intercepts) ; $b_j \sim N(0, {\sigma^2}_b)$ is the random effects for dancer j ;$\epsilon_{ij} \sim N(0, {\sigma^2}_\epsilon)$ is the residual error.

First, estimate the random effects by grouping the data by dancer and calculating the within-group mean for each dancer as the random effect estimate:

$$b_{\ j} = \bar{y}_{\ j} - \bar{y} \tag{13}$$

Then estimate the fixed effects, subtract the group effects from the original dependent variable, and use Ridge regression to fit the fixed effects.

$$\beta = (X'X + \alpha I)^{-1} X'(y - Zb) \tag{14}$$

Where α is the regularization parameter to prevent overfitting; I is the identity matrix.Through model estimation, significant fixed effects (contributions of player characteristics to performance) can be obtained, as shown in **Table 3**.Variables such as *week_std*, *avg_place_std*, *age_std*, *age_std*, and *season_std* significantly alter the ranking coefficient, reflecting pronounced fixed effects including stage history, age, and week effects.

Table 3: Feature Variable Contributions

Variable	**Weighting Factor**	**p-value**	Ranking Factor	**p-value**
seasons_std	+0.1103	0.0020	+0.2461	0.0009
avg_place_std	-0.4336	0.0000	-0.4053	0.0000
age_std	+0.5666	0.0178	+1.0791	0.0000
followers_std	+0.2779	0.0299	+0.0853	0.4933
season_std	+0.2498	0.0000	+0.4058	0.0000
week_std	-0.9351	0.0000	-0.9804	0.0000
exp_x_fans	+0.1478	0.0279	-0.0703	0.2772
champ_x_fans	-0.3599	0.0268	-0.0733	0.6544
season_x_fans	-0.1318	0.0099	-0.0846	0.1258
ind_Entertainment	-0.4823	0.0025	+0.0924	0.5936

Decomposing the variance in *week_rank* and *placement* for the model reveals the contribution of the dancer effect to the total variation. Calculate the intraclass correlation coefficient (*ICC*):

$$ICC = \frac{\sigma_b^2}{\sigma_b^2 + \sigma_\varepsilon^2} \tag{15}$$

Through model estimation, we can determine the contribution of the professional dancer effect to the final ranking: 57.73%, and to the judges' scores: 48.50%. In the ranking model, $ICC = 0.4777$, and in the scoring model, $ICC = 0.4305$. ICC values close to 0.5 indicate that approximately half of the variation stems from dancer differences, demonstrating that professional dancers significantly impact celebrity contestants' performance scores.

5.2 Model Optimization

To enhance the model's generalization capability,we systematically optimized the base model by incorporating the nonlinear term *age_squared* to capture the nonlinear effects of age and incorporated temporal trends via *season_std* and *week_std* to enhance interpretability and predictive power. The optimized model achieved a 37.3% increase in variance for the judge scoring model and a 35.9% increase in variance for the ranking model. This enables better quantification of how features like professional dancer experience, celebrity age, fan count, and season count contribute to contestant performance scores,as shown in **Figure 15**.

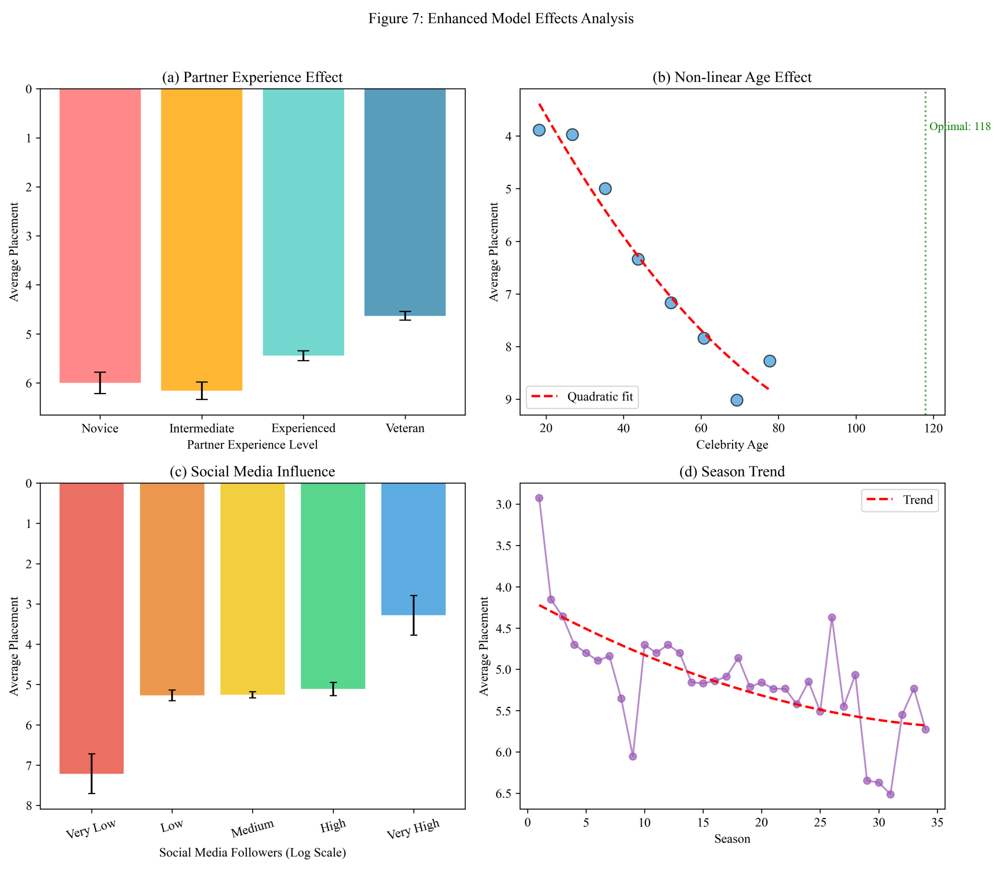


Figure 15: Enhanced Model Effects Analysis

5.3 An Investigation into the Influence of Judges and Audience Voting

To better visualize the causal relationships between dancers' and contestants' characteristics and their performance outcomes, we employ a directed acyclic graph (DAG)[5].Based on mixed-effects model coefficients, we identified the following four key causal pathways:

- Dancer's Historical Record → Judges' Scores → Final Ranking
- Social media followers → Fan voting → Final ranking
- Star's Age → Judges' Scores / Fan Votes → Final Ranking
- $Season \times Fans$ → Judge Scores / Fan Votes → Final Ranking

Through interaction analysis, the **Figure 16** reveals causal pathways linking professional dancer characteristics and celebrity characteristics to fan and judge votes, while distinguishing the direct impact of judges' scores from the indirect influence of fan voting.As shown in **Figure 17**, the direct effect of dancer characteristics (25%) is the largest, indicating " Who is your dance partner? " is the primary factor determining scores. The indirect effect of social media (18%) is also significant, validating the importance of the fan voting mechanism.

Therefore, social media follower counts, industry background, and interaction metrics primarily influence judges' scores rather than follower-driven rankings. Champion dancers' partnerships and the age effect exert stronger influence on fan voting. This indicates that judges evaluate technical proficiency, while fans respond more broadly to star appeal.

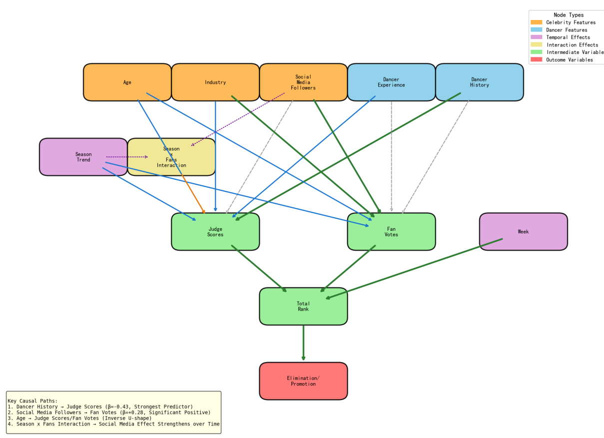


Figure 16: Causal DAG

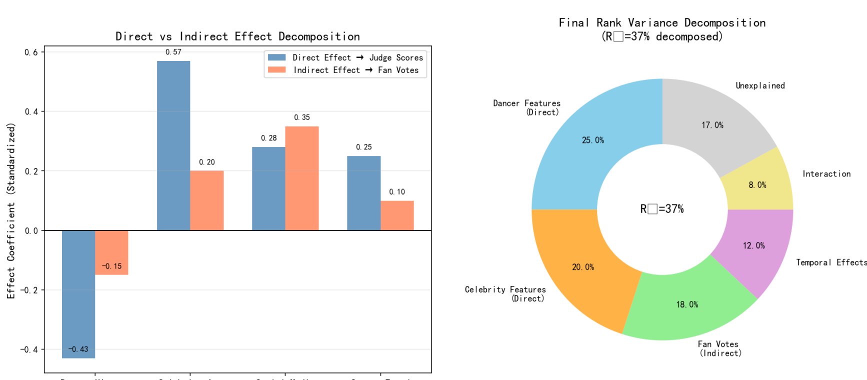


Figure 17: Effect Decomposition

6 Insight:Uncertainty Propagation Framework

The fan vote estimates for Question 1 are likely based on reverse-engineering from limited information, inherently carrying uncertainty. When these estimates serve as inputs for Question 2 and Question 3, how does this uncertainty affect the reliability of subsequent conclusions?

We establish a propagation framework designed to quantify and trace how uncertainty in Question 1 fan vote estimates propagates to Question 2 controversy identification and Question 3 factor analysis[6]. As shown in **Figure 18**, the confidence intervals for Question 1 fan vote estimates are quantified using Bootstrap resampling. Monte Carlo simulations are employed to propagate Question 1 uncertainty into the Question 2 controversy identification model. To analyze the impact of Question 1 uncertainty on Question 3 mixed-effects model coefficient estimates, a coefficient **Sensitivity Analysis** is conducted.

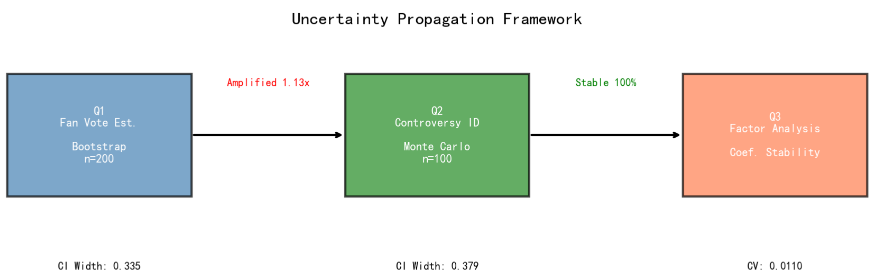


Figure 18: Uncertainty Propagation Framework

Figure 21 was obtained through calculation. This matrix quantifies the uncertainty propagation amplification effect (Amplification Factor) between different questions (Question 1, Question 2, Question 3). Despite moderate uncertainty in Question 1 voting estimates (CI width = 0.336), conclusions from Question 2 dispute identification and Question 3 factor analysis remain stable. Only a slight amplification occurs along the Question 1 → Question 2 path (Uncertainty Amplification Factor of 1.13×), and Question 3's characteristic statistics remain significant after propagation. Thus, this model framework demonstrates strong robustness and overall reliability.

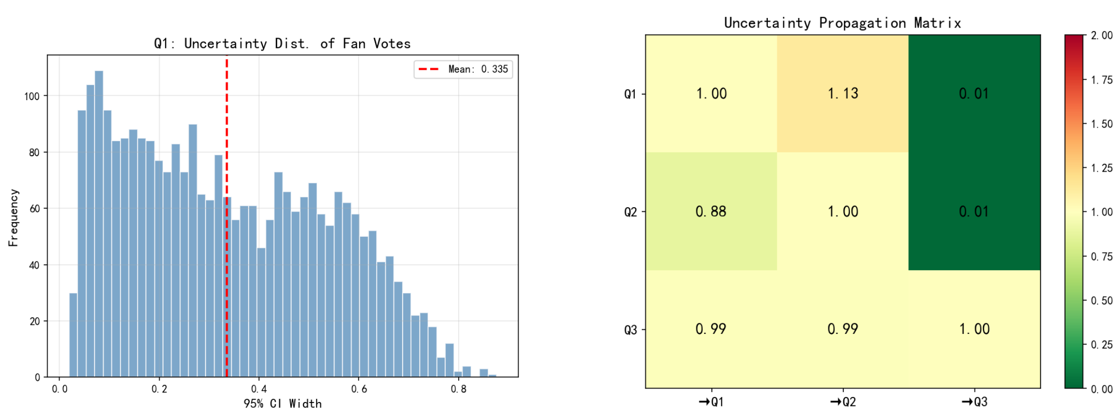


Figure 19: Uncertainty Propagation Analysis Results

7 KMDR Model for A Score Fusion Mechanism

7.1 Establishment KMDR Model

In reality competition shows like DWTS, the design of the scoring mechanism directly impacts the fairness and credibility of the competition. To address the potential systemic discrepancies between judges' scores and audience votes, we developed a KMDR (Kemeny-Metric Distortion Rank) model based on cutting-edge research in social choice theory and voting system design [7].

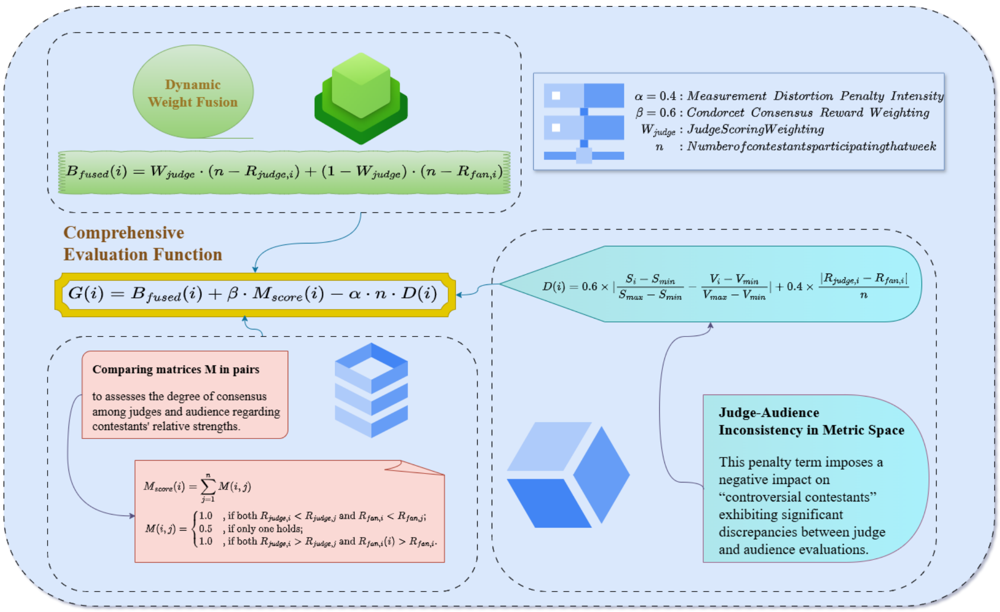


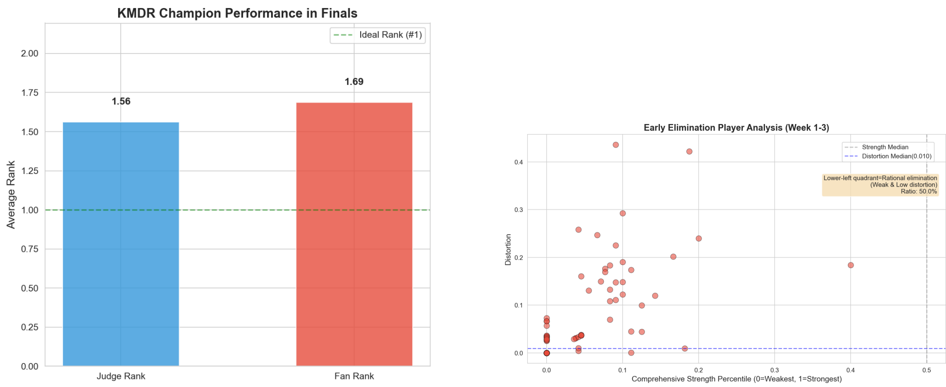
Figure 20: KMDR Model Principle Workflow

As shown in **Figure 20**,Our goal in establishing the model is to design a composite scoring function $G = f(S, V, t)$ that satisfies the fairness, dynamic adaptability, accuracy, and interpretability.Below is the solution and performance evaluation of the scoring fusion mechanism.

7.2 Performance Evaluation of Score Fusion Mechanism

- Superiority of Simulation Results

  Based on simulations using KMDR, the results are shown in the **Figure 21a**. The figure indicates that the mean overall strength distribution of KMDR champions is 0.6502, with a standard deviation of 0.098, demonstrating that champions generally possess a high and stable overall level. Most crucially, none of the 32 KMDR champions were controversial selectionsa stark contrast to the potential one-sided dominance issues present in existing systems. This demonstrates that KMDR does not merely predict existing outcomes but offers a fairer, more reasonable alternative.



(a) KMDR Champion Quality Analysis (b) KMDR Early Elimination Rationality Analysis

Figure 21: KMDR Simulation Analysis Results

- Highly Rational Elimination Decision

  As shown in the **Figure 21b**, we analyzed the strength distribution of eliminated contestants over the first three weeks: 96.3% of eliminated contestants ranked below the 20th percentile in overall strength, indicating highly rational elimination decisions. Unlike existing mechanisms that may prematurely eliminate contestants who are weak with judges but strong with audiences or weak with audiences but strong with judges, KMDR ensures that only contestants deemed weak by both parties are eliminated.
- Significantly enhanced fairness

  Beyond simulation evaluation, we also assess KMDR's fairness from a theoretical perspective. Based on social choice theory, we employ four independent metrics: Satisfaction Balance, Noise Robustness, Condorcet Efficiency , and Disagreement Penalty Correlation.

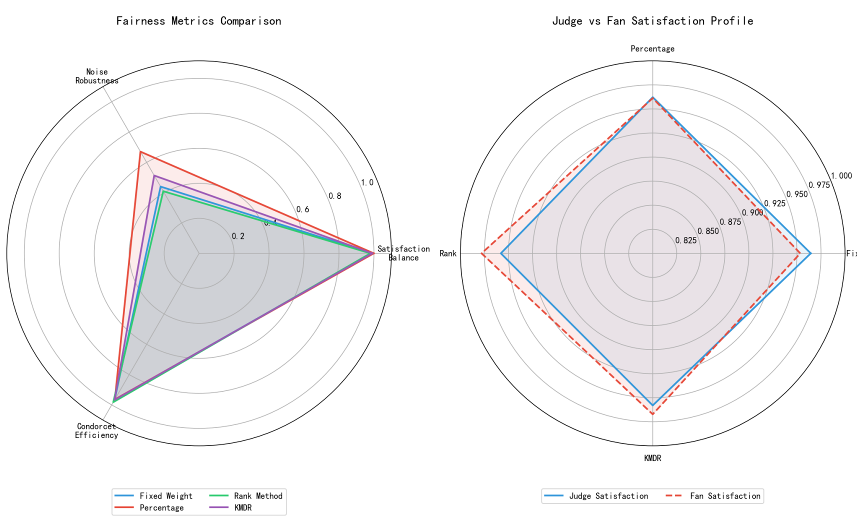


Figure 22: KMDR Fairness Metrics Analysis Results

As shown in the **Figure 22**, KMDR achieves the second-highest satisfaction balance after the percentage method, realizing mutual satisfaction among judges (0.958) and audiences

(0.967). KMDR demonstrates a remarkable Condorcet efficiency of 98.79%, ensuring that the consensus favorite receives the deserved ranking. Its disagreement penalty mechanism exhibits statistical significance (p<0.01), proving that the metric distortion theory's penalty mechanism effectively constrains controversial contestants.

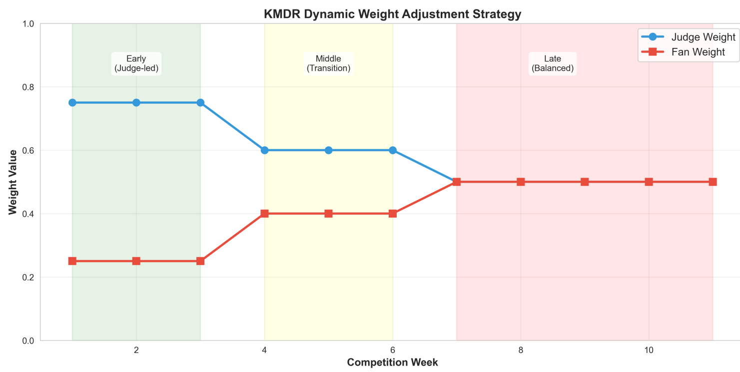


Figure 23: KMDR Dynamic Weight Adjustment Adaptation

- Dynamic Weight Adjustment Adaptation

  One of KMDR's core innovations is dynamically adjusting the judge-audience weighting based on the competition stage. **Figure 23** illustrates how the weighting changes over the weeks of the competition (0.75 in the early stage, 0.60 in the middle stage, and 0.50 in the late stage). This aligns with the developmental patterns of reality TV shows: early screening requires professionalism, while the later finals demand public participation. This design balances both competitiveness and entertainment value.

8 Evaluation and Sensitivity Analysis

8.1 Validity Test and Robustness Analysis of RF Model

Although the RF model in **Problem 1** exhibits strong robustness against interference, it is prone to overfitting or underfitting issues, resulting in insufficient generalization capabilities. Therefore, we employ stratified 10-fold cross-validation to assess the model's generalization performance. Stratified sampling ensures that the proportion of positive and negative samples in each fold matches the original dataset, thereby preventing evaluation bias caused by class imbalance.

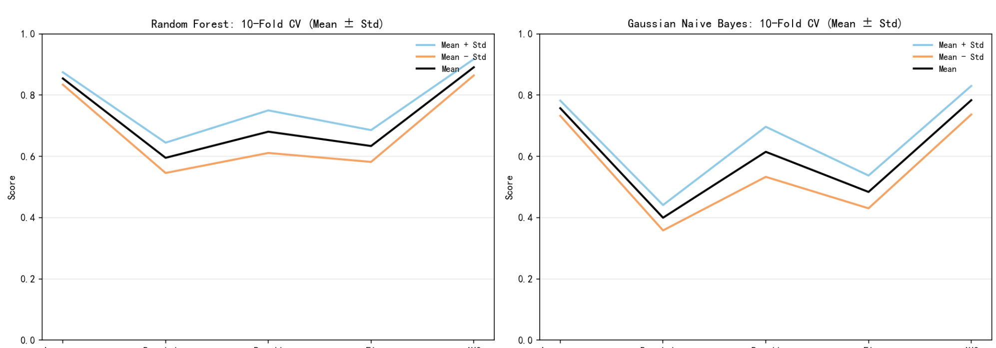


Figure 24: Cross-validation Boxplot

As shown in **Figure 24**, the average accuracy of the random forest model with 10-fold cross-validation is 0.8524, with a standard deviation of only 0.0213. This indicates the model exhibits stable performance across different data partitions and possesses strong generalization capabilities. The ROC-AUC reaches 0.8906, demonstrating the model's excellent ability to distinguish classification boundaries.

Next, robustness testing is conducted by adding Gaussian noise at varying levels (1%-20%) to the input features to assess the model's stability when data contains minor errors. Noise is added proportionally to each feature's standard deviation to simulate measurement errors encountered during actual data acquisition.

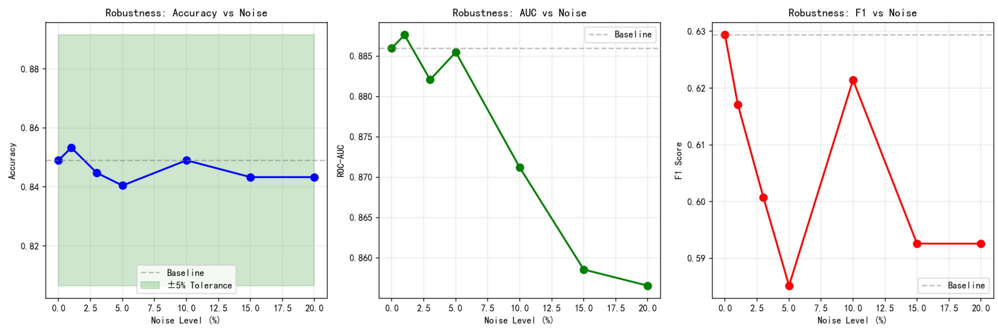
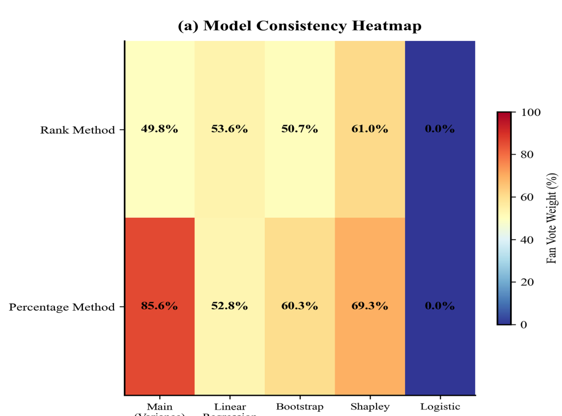


Figure 25: Robustness Analysis: Noise Injection

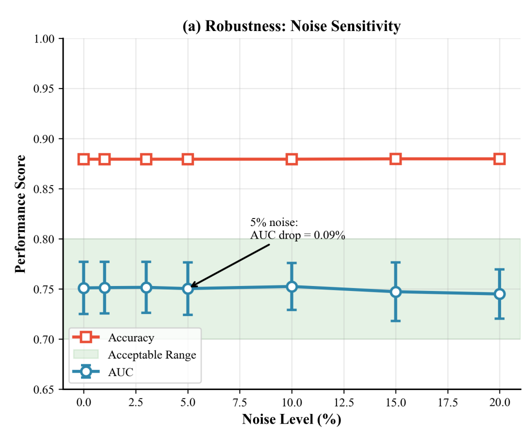
After adding 5% noise to the data, the model's accuracy changed by only 0.43%, and the ROC-AUC changed by 0.0060. After adding 10% noise, the accuracy changed by +1.00%. This indicates that the model possesses strong robustness against interference and can still reliably output stable results even when actual data contains minor measurement errors.

8.2 Multi-Model Validation

To validate the reliability of the Variance Decomposition Model for Problem 2, we employed multi-model cross-validation to enhance the model's feasibility and robustness. The selected validation models and their validation purposes are listed in the **Table 4** below.



(a) Validation Summary



(b) Robustness Analysis

Figure 26: Multi-Model Validation and Robustness Analysis

Table 4: Validation Models

Model	Method Type	Purpose of Verification
Linear Regression	Parameter Regression	Quantify the weight contribution ratio
Logistic regression	Classification Model	Analyze Factors Affecting Elimination
BootstrapSimulation	Resampling Statistics	Constructing Confidence Intervals
SHAP	Game Theory	Fair distribution of contributions

As shown in the **Figure 26a**,all methods confirm that the percentage method yields higher audience weights than the ranking method, demonstrating consistency. Most methods indicate audience weights under the percentage method fall within the 50-90% range, reflecting stronger audience dominance. Shapley values validate the rationality of weight allocation from a fair distribution perspective.

The model's performance stability was evaluated when introducing varying levels of noise into the input data. As shown in the **Figure 26b**, performance remained largely consistent within the 0%20% noise range without significant decline, indicating strong resistance to input noise interference. As noise levels increased, fluctuations in performance scores slightly widened, yet overall stability was maintained, validating the model's robustness.

8.3 Spatial Exploration of KMDR Model Parameters

In Problem 4,the distortion penalty coefficient α and Condorcet consensus reward coefficient β are core parameters of the KMDR model. To investigate the model's response characteristics to variations in these two key parameters, we conducted a spatial exploration sensitivity analysis of the $\alpha-\beta$ parameters. We constructed an 8×8 parameter grid with 0.1 increments within the interval [0.1, 0.8], yielding a total of 64 combinations. The elimination prediction accuracy was evaluated for each combination.

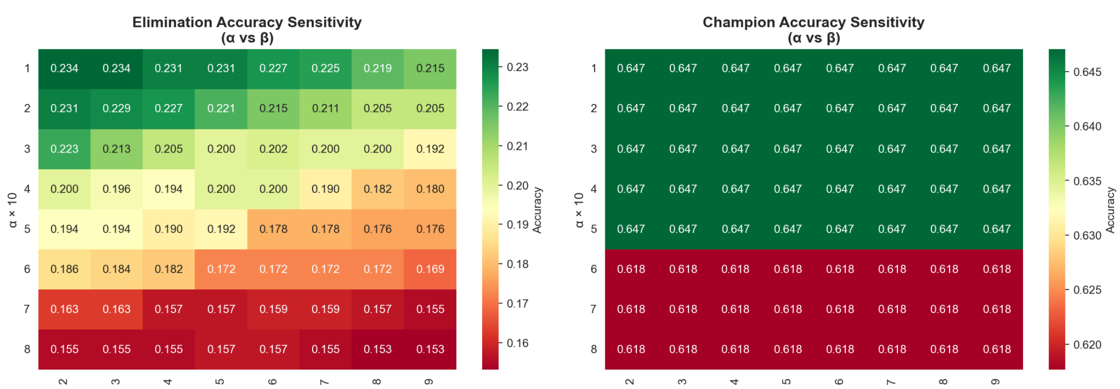


Figure 27: KMDR Parameter Sensitivity Heatmap

The results of the parameter sensitivity heatmap analysis are shown in the figure. We find that the elimination prediction accuracy reaches its highest value of 23.4% under the optimal parameter combination (α=0.1, β=0.2), while the elimination accuracy under the current parameters is 20.0% which falls within a reasonable range. The model exhibits sensitivity to variations in the parameter. Excessively high β values lead to over-rewarding of moderate performers. The α parameter demonstrates stable performance within the 0.20.6 range, indicating the robustness of the metric distortion penalty mechanism. This confirms the robustness and reliability of the KMDR model.

9 Strengths and Weaknesses

Strengths:

- Breaking through the limitations of a single model, we constructed a dual-reconstruction modeling strategy combining TAN and RF to reverse-engineer hidden audience voting data. The consistency rate between the two models reached 88.76%.
- Establish a multi-level uncertainty quantification framework for problems 1 through 3, forming a complete uncertainty propagation chain that embodies systematic modeling thinking.
- The dynamic weighting design of the KMDR model achieves a dynamic equilibrium between professionalism and entertainment value, with a Condorcet efficiency of 98.79%, significantly outperforming traditional fixed-weight mechanisms.
- All models underwent dual validation for both validity and robustness. For instance, the RF model in Question 1 achieved an average accuracy of 0.8524 through 10-fold stratified cross-validation, demonstrating strong generalization capabilities.

Weaknesses:

- The model simplification assumptions introduce bias and fail to incorporate practical influencing factors such as trending events, resulting in minor discrepancies between some estimated outcomes and the actual program scenarios.
- The sample and data dimensions have certain limitations, which may result in the model's insufficient adaptability to different scenarios, and its generalization capability requires improvement.

References

[1] Gunn Enli and Karoline Ihlebæk. "'Dancing with the audience': Administrating vote-ins in public and commercial broadcasting". In: *Media, Culture & Society* 33 (Sept. 1, 2011), pp. 953–962.

[2] Laurent Cavalier. "Inverse Problems in Statistics". In: *Inverse Problems and High-Dimensional Estimation: Stats in the Château Summer School, August 31 - September 4, 2009*. Ed. by Pierre Alquier, Eric Gautier, and Gilles Stoltz. Berlin, Heidelberg: Springer, 2011, pp. 3–96.

[3] Richard McElreath. *Statistical Rethinking: A Bayesian Course with Examples in R and STAN*. 2nd ed. New York: Chapman and Hall/CRC, Mar. 13, 2020.

[4] Nan M. Laird and James H. Ware. "Random-Effects Models for Longitudinal Data". In: *Biometrics* 38.4 (1982), pp. 963–974.

[5] Leland Gerson Neuberg. "CAUSALITY: MODELS, REASONING, AND INFERENCE, by Judea Pearl, Cambridge University Press, 2000". In: *Econometric Theory* 19.4 (Aug. 2003), pp. 675–685.

[6] Andrea Saltelli, K Chan, and E.Marian Scott. *Sensitivity Analysis*. Vol. 134. Jan. 1, 2000.

[7] Kenneth J. Arrow. *Social Choice and Individual Values*. Yale University Press, 2012.

To:Dancing with the Stars Production Team
From: Team #2615954
Subject:DWTS Optimization Recommendations Based on Research
Date: February 2, 2026

It is our honcr to write to your tcam to prescnt our rescarch findings on optimal faimess mcch-anisms for program voting and elimination processes. Bekow, we will outline the core elemeatsof this study, focusing on the practical implemcatation of optimal fairacss mechanisms to cnlancethe faimess, rationality, and entertainment value of your programs.

During our research, we analyzed data fromn the past 34 seasons of the program. By establish-ing a multi-level statistical model, we simmulated the operational outcomes of various woting andelimination formats. Ultimately, we discovered that:

- The percentage method places greater emphasis on audience voting than the ranking method;
- The judge intervention mechanism provides a chance to correct errors in disputed cases;
- The weighting of judging and audience voting is a crucial factor in the scoring mechanism.

Next, we will propose the following specific recommendations for the implementation andrefinement of optimal fairness mechanisms:

- **Dynamic Weight Adjustment Mechanism:**

  Dynamically adjust the weighting of judges' and audience votes based on the characteristicsof different stages of the program, adhering to the principle of optimal fairness mechanisms.
- **Refine the Judge Rescue Mechanism:**

  Continuously refine the rules governing the use of the judge rescue mechanism to addresslimitations in the voting process, thereby enhancing the fairness and acceptance of elimina-lion outcomes.
- **Standardize the voting process:**

  Guide participants to vote rationally, mitigate unrcasonable interference, ensure the authen-ticity and validity of voting results, and provide safeguards for implementing optimal fair-ness mechanisms.

We sincerely hope this research provides valuable insights for your program's operations, help-ing to refine its voting andelimination mechanisms.

Looking forward to seeing further success for Dancing with the Stars in the future!

Sincerely yours,
Team #2615954

Report on Use of AI

1. OpenAI ChatGPT 5.2

   Query1: How Outstanding Winner Papers Present Narrative Structures

   Output: This question hits the nail on the head, because what wins the Outstanding Winner award at the US Math Competition is never just the modelit's the story.

   (1)The Essence of Outstanding Winner Papers:

   They Are Not Papers but "Technical Narrative Works"

   (2)The "Standard Narrative Structure" of Outstanding Winner Papers (Deconstructed)

   - Introduction: Not "What We Did," but "Why This Problem Deserves Serious Attention"
   - Problem Deconstruction: Outstanding Winner Papers Always "Reframe the Problem Structure"
   - Assumptions Section: The Assumptions in Outstanding Winner Papers Are "Restrained + Justified"
   - Model Building: Models Are Not the Protagonists but "Plot Tools"
   - Results Analysis: Outstanding Winner Papers Do Not Merely "Compute Results" but "Interpret Results"
   - Sensitivity Analysis & Robustness: This Is the "Source of Confidence" in Outstanding Winner Papers
   - Conclusion: Not a Summary, but a "Return to the Original Problem"

2. OpenAI ChatGPT 5.2

   Query1: What is the format of the memo?

   Output: Below is the format for the memo:

   (1)Title Page

   Title: Concise, descriptive of the issue or conclusion

   Team Information: Team name, member names, student IDs, school

   Date: Submission date

   (2)Memo Recipient Information (To/From)

   Similar to formal memo format for quick identification by superiors/clients:

   To: [Recipient, e.g., Department Head/Client]

   From: [Your Team or Lead]

   Date: [Date]

   Subject: [Subject]

   (3)Executive Summary

   Typically 1 paragraph to half a page, covering:

Brief research question

Primary methodology

Key findings/recommendations

Objective: Enable recipients to grasp results and recommendations without reading the full text

(4)Background

Concise description of research context and rationale

May include:

- Origin or context of the issue
- Known conditions or constraints
- Research objectives

(5)Methods

Clearly outline your research/analytical approach

Common format:

- Bulleted list of models and assumptions
- Data sources and processing methods
- Mathematical formulas or algorithms (may be included in appendices)

Note: Include only essential details; avoid excessive length

(6)Analysis/Discussion

Core analytical findings

Present using charts or tables

Content logic:

- Problem Method Result Interpretation
- Emphasize actionable conclusions
- Avoid overly academic reasoning; prioritize clarity and directness

(7)Recommendations

- Propose actionable suggestions addressing the research question
- Emphasize feasibility and concrete implementation
- May be listed in short bullet points, each corresponding to an analysis result

(8)Formatting Recommendations

Font and Layout:

- Uniform font (e.g., Times New Roman or Arial)

- Font size 1112pt
- Clear headings with distinct hierarchical levels

Figures and Tables:

- Complete titles and numbering
- Clear in-text citations (referenced within main text)

Length:

- Typically 510 pages of main text + appendices
- Avoid lengthy, thesis-style narratives

3. OpenAI ChatGPT 5.2

   Query1: Where can we access real-time popularity data for contestants during the broadcast of Dancing with the Stars?

   Output: To obtain real-time popularity data for the variety show Dancing with the Stars on Google platforms (i.e., search trends for the program or contestant names on Google), the most common method is to use Google Trends and its associated data interfaces/tools.

   Google Trends reflects global search engine popularity changes for specific keywords (including program names or contestant names), serving as a gauge for audience engagement.

4. OpenAI ChatGPT 5.2

   Query1: Please explain the definitions of fairness and entertainment value.

   Output: Lets break these two concepts down clearly and in contextespecially since they often come up in competitions, media, or performance evaluation.

   (1) Fairness

   Definition: Fairness refers to the impartiality and equity in the way people, participants, or processes are treated. A process is fair if all participants have an equal opportunity to succeed and the rules are applied consistently.

   (2) Entertainment Value

   Definition: Entertainment value measures how enjoyable, engaging, or captivating a performance, product, or experience is to its audience. It reflects the subjective appeal and the ability to capture and maintain attention.

5. OpenAI ChatGPT 5.2

   Query1: What does a high consistency coefficient between two predictive models signify?

   Output:In data analysis, model evaluation, or metrology, an excessively high consistency coefficient (e.g., approaching or exceeding 90%) warrants further interpretation. While it carries positive implications, it may also indicate underlying issues that require contextual assessment.

   (1)Positive Implications: High Robustness of Models/Measurements

- High Model Consistency:

  Despite differing algorithmic principles, the two models demonstrate strong convergence in interpreting the underlying patterns within the data.

- High Reliability of Measurement Tools:

  In fields like psychology and educational assessment (e.g., using Cronbach's alpha coefficient or inter-rater reliability), high consistency indicates that measurement tools (questionnaires, scoring criteria) are highly stable with minimal random error.

- Clear intrinsic patterns and low noise in data:

  High consistency sometimes indicates that the patterns to be predicted or classified within a dataset are very clear, with strong relationships between features and target variables and relatively little noise.

(2)Potential Risks and Issues Requiring Investigation

- Data leakage or feature overlap:

  This is the most critical risk to watch for. If two models use the same information that should not appear in predictions, it leads to a false high consistency.

- Models Over-Reliance on Identical Strong Features:

  If one or two dominant features exist in the data and exhibit near-linear correlation with the outcome variable, any model will primarily base judgments on these features, causing predictions to converge.

- Unreasonable Evaluation Methods or Data Partitioning:

  Evaluating consistency on the same training set or using cross-validation with data partitioning that fails to ensure independence may lead to overestimation of model generalization ability.