

Problem Chosen

C

2026

**MCM/ICM
Summary Sheet**

Team Control Number

2400860

Summary

In the operation of the renowned entertainment competition show Dancing with the Stars (DWTS), significant discrepancies between professional judges' scores and audience voting preferences have exposed flaws in the existing elimination mechanism. This paper, grounded in multi-statistical modeling principles, primarily investigates big data analysis and optimization of the voting mechanism.

For Problem 1, we employed a dual reconstruction modeling strategy combining the TAN model with the Random Forest model and SHAP interpretability analysis to estimate hidden audience voting data and quantify uncertainty. The results showed that the TAN model achieved an accuracy of 82.48%, while the RF model reached 84.90%. The consistency in prediction direction between the two models was 88.76%, enabling precise reconstruction of hidden audience voting data and quantification of its uncertainty.

For Problem 2, we combined counterfactual simulation and bias sensitivity analysis methods to construct a variance decomposition model based on audience voting data from Problem 1. By setting thresholds for audience vote weight calculations and validating the rationality of two vote synthesis mechanisms through Spearman's correlation analysis, we determined that the percentage method (85.6%) significantly outweighs the ranking method (49.8%) in audience vote weighting, demonstrating greater preference for audience votes. The judge rescue mechanism can correct 8.9% of controversial elimination outcomes, which can significantly enhanced the fairness of the program when combined with the ranking method.

For Problem 3, we constructed a mixed-effects model incorporating ridge regression regularization and calculated intraclass correlation coefficients (ICC) to quantify the contribution of factors such as professional dancers and contestant characteristics to elimination outcomes. Additionally, we employed directed acyclic graph (DAG) to uncover causal relationships among variables. Results indicate that professional dancers significantly influence contestant rankings (57.73%) and judge scores (48.50%). Variables like weekly contestant scores and age exert greater weight on outcomes, while fan following indirectly impacts final rankings by influencing audience votes.

For Problem 4, through analysis of dynamic weight fusion, contestant consensus scoring, and metered distortion penalty mechanisms, we established a score fusion mechanism based on the KMDR model. By optimizing weight parameters via $\alpha - \beta$ parameter grid search and setting four fairness evaluation metrics including Condorcet efficiency and satisfaction balance we designed an optimal voting scoring mechanism. Simulation results demonstrate a 100% undisputed champion rate and 96.3% early elimination rationality. The dynamic weight parameter adjustment strategy achieves a balanced blend of professionalism and entertainment value.

We also constructed an uncertainty propagation framework to analyze the transmission of uncertainty from voting data estimation through mechanism rationality analysis to the quantification of eigenvalue contributions. Finally, we conducted robustness and parameter sensitivity analyses on the model while evaluating its strengths and weaknesses.

Keywords: TAN ; Random Forest ; Mixed-Effects ; Voting Mechanism

Contents

1	Introduction	3
1.1	Problem Background	3
1.2	Restatement of the Problem	3
1.3	Our work	3
2	Preparation for Models	4
2.1	Assumptions and Notations	4
2.2	Data Processing	5
3	Fan Vote Estimation Based on TAN and SHAP Models	6
3.1	Tree Augmented Naive Bayes Model for Fan Vote Estimation	6
3.2	RF and SHAP Model for Fan Vote Estimation	8
3.3	Consistency Test of Two Models	10
3.4	Validity Test and Robustness Analysis	11
4	Voting Combinations Comparison via Variance Decomposition Model	12
4.1	Counterfactual Simulation Modeling	12
4.2	Bias Sensitivity Analysis	13
4.3	Analysis of Controversial Cases and Method Evaluation	14
4.4	Discussion on the Judge Rescue Mechanism	15
4.5	Multi-Model Validation and Robustness Analysis	17
5	Quantifying Feature Contributions via Mixed-Effects Model	18
5.1	Calculate Fixed and Random Effects	18
5.2	Model Optimization	19
5.3	An Investigation into the Influence of Judges and Audience Voting	20
6	Insight:Uncertainty Propagation Framework	21
7	Design of a Score Fusion Mechanism Based on the KMDR Model	22
7.1	Establishment and Evaluation of the Score Fusion Mechanism	22
7.2	Sensitivity Analysis	25
8	Sensitivity and Robustness Analysis	25
9	Strengths and Weaknesses	25
	Appendices	28

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 competition and even win by 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 judges' bottom two elimination 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.

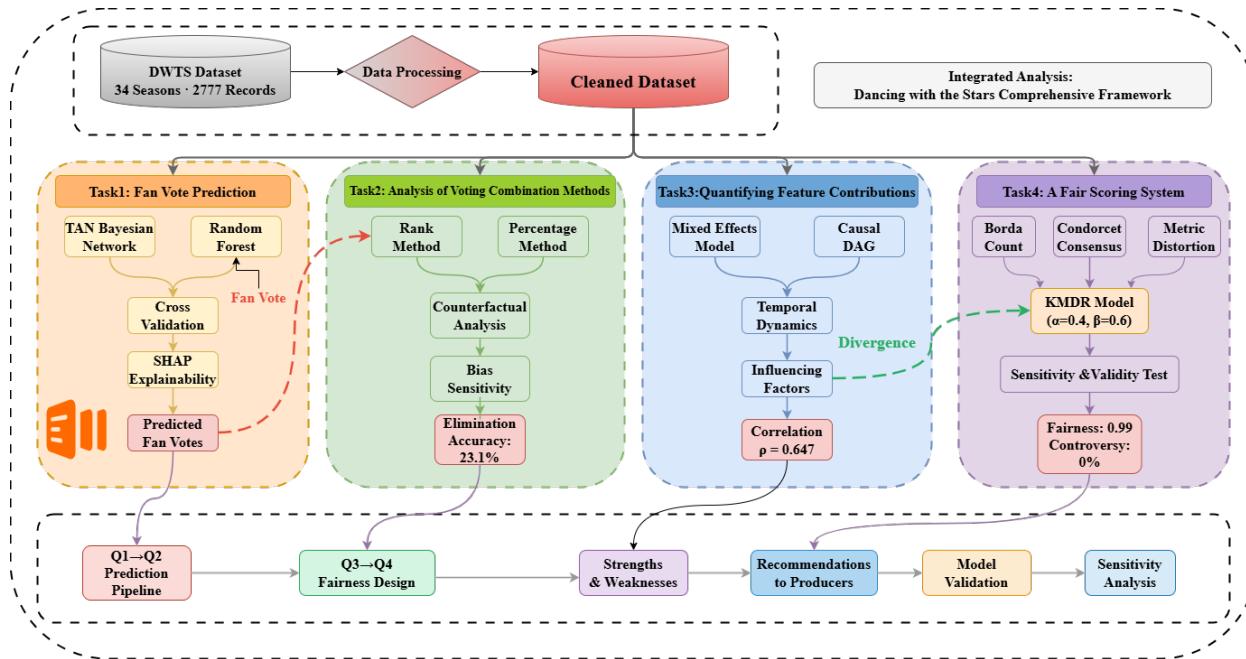


Figure 1: Our work

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:** The data we use are accurate and valid.
- **Justification:** Our data is collected from the World Bank and some other official web sites and research papers. its reasonable to assume the high quality of their data.
- **Assumption:** The regions under study will remain peaceful and stable, with no significant events other than natural disasters occurring in the foreseeable future.
- **Justification:** A stable capital market environment provides a predictable framework within which we can project our expected returns. It is important to note that this assumption does not negate the potential impact of natural disasters.
- **Assumption:** The estimated figures for each region represent an average level of performance or condition for that area.
- **Justification:** For the purposes of this study, treating each region as a cohesive entity allows for a more streamlined analysis. This methodological approach simplifies the complexity inherent in regional studies by focusing on aggregate data, thereby providing a generalized view of each area's performance or condition.

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

Table 1: Notations

Symbol	Definition
V_i	Fans Vote Score of the i -th contestant
Y	$Y \in \{0, 1\}$:Elimination Flag (0=Not Eliminated, 1=Eliminated)
X	$X = (X_1, X_2, \dots, X_n)$:Feature Vtor
S_i	Judge score of the i -th contestant
R_i	The rank of the i -th contestant

* 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.

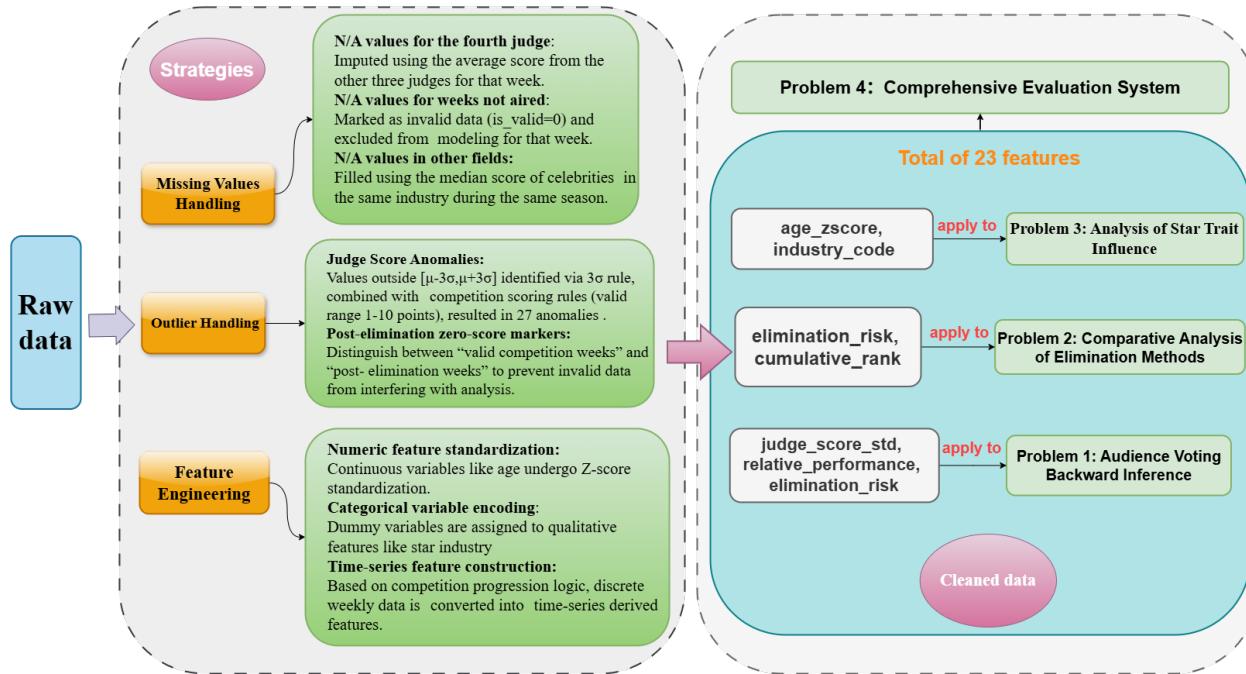


Figure 2: Data Processing Workflow

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

In the show Dancing with the Stars, a common scenario occurs: some celebrities receive low scores from professional judges yet advance smoothly or even win the championship thanks to high audience support rates. 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 real-world data, features often exhibit dependencies. 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:

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

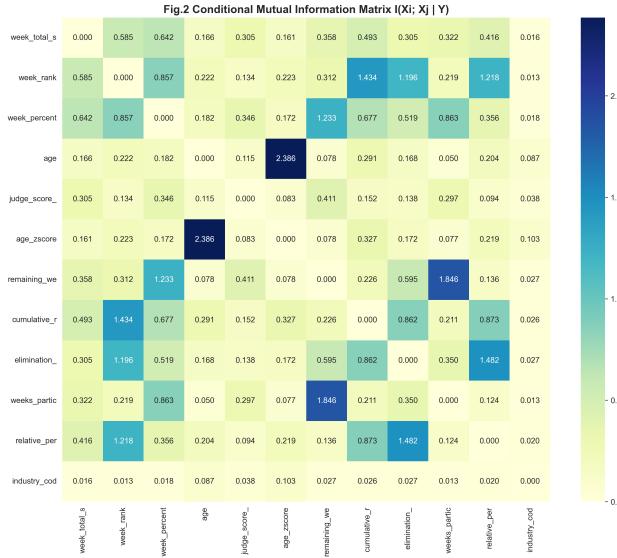


Figure 3: Conditional Mutual Information Matrix

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

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

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 $X = (X_1, X_2, \dots, X_n)$, compute the posterior probability via Softmax normalization:

$$P(Y = 1 | X) = \frac{\exp(\ell_1)}{\exp(\ell_0) + \exp(\ell_1)}, \quad (3)$$

$$\ell_j = \log \tilde{P}_j \quad (j = 0, 1)$$

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

$$V = (1 - P((Y = 1) | X)) \times 100 \quad (4)$$

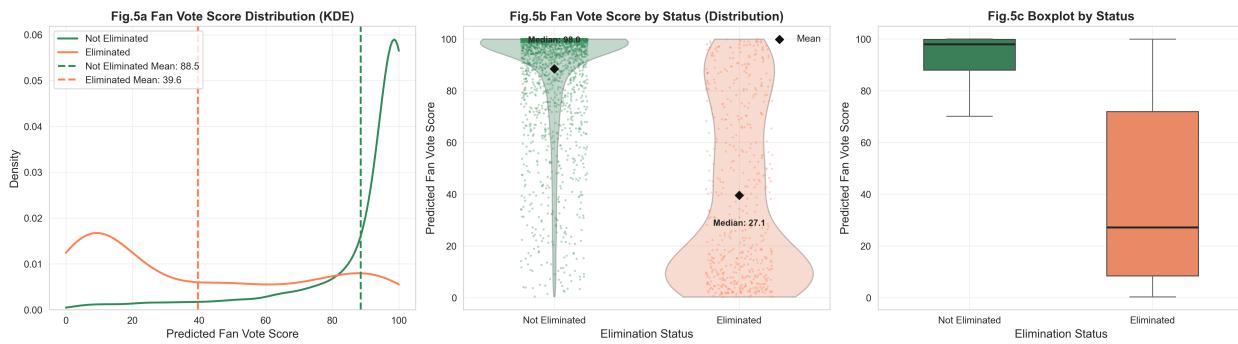


Figure 4: Fan Vote Score Distribution

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.

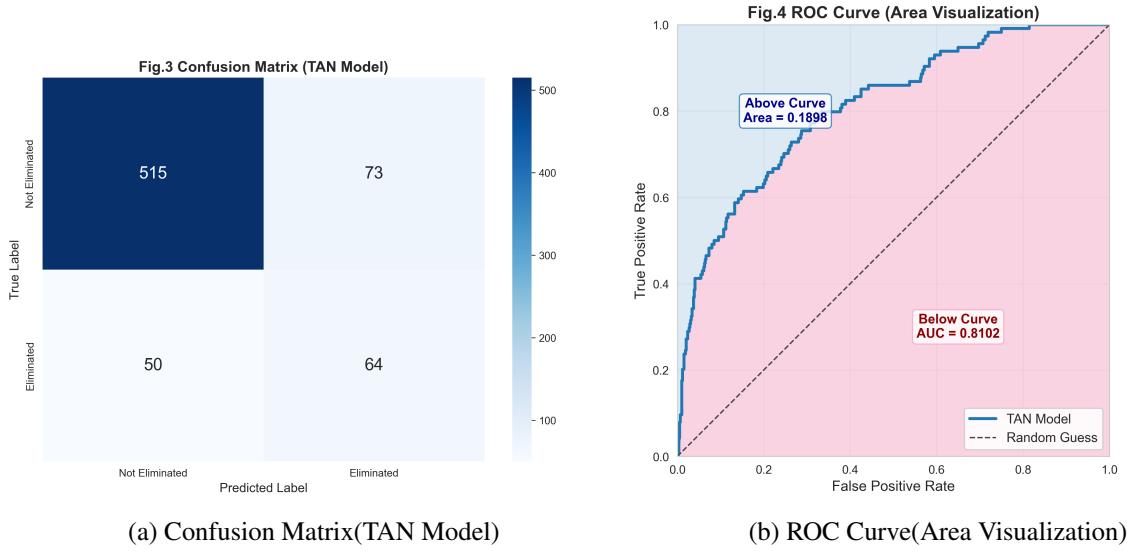


Figure 5: Confusion Matrix and ROC Curve

3.2 RF and SHAP Model for Fan Vote Estimation

The Random Forest classifier, combined with SHAP (SHapley Additive exPlanations) interpretability analysis, infers audience voting scores from elimination results. The Random Forest prediction formula is:

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

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:

$$P((Y = 1) | x) = (1/K) \times \sum_k I(h_k(x) = 1) \quad (6)$$

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.

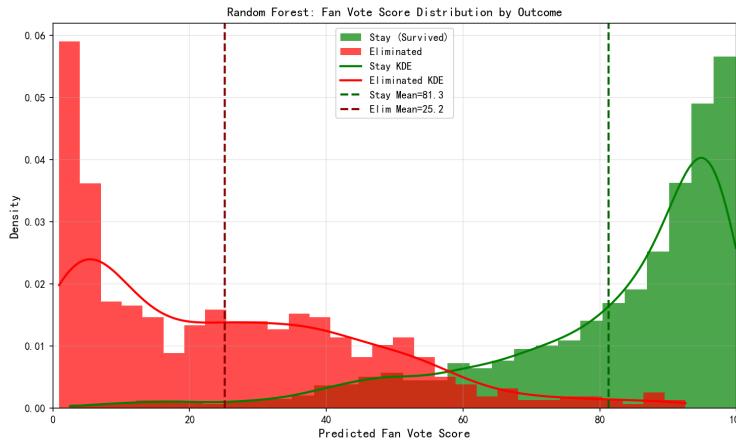


Figure 6: Fan Vote Score Distribution(Random Forest Model)

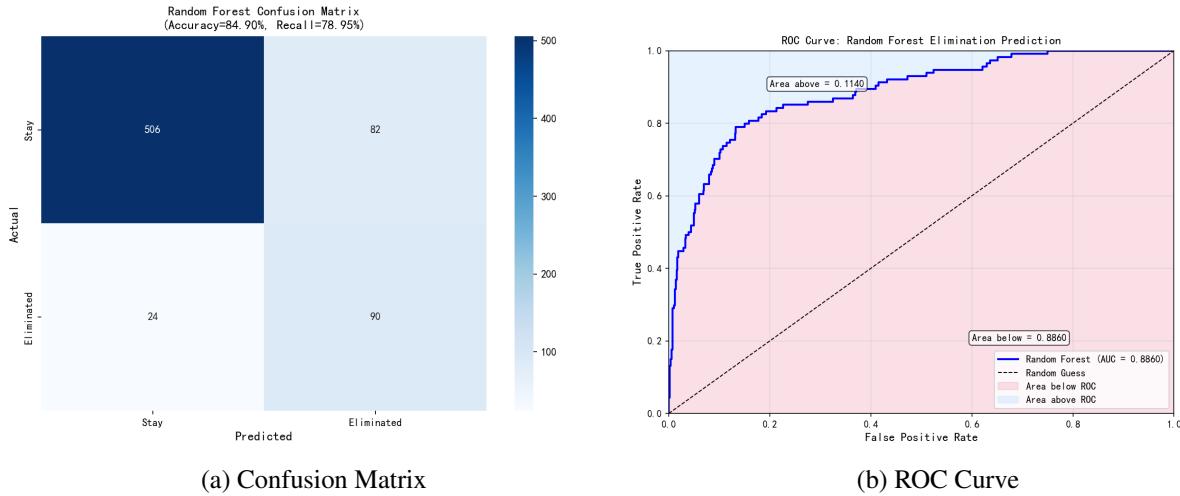


Figure 7: RF Model Confusion Matrix and ROC Curve

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:

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

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.

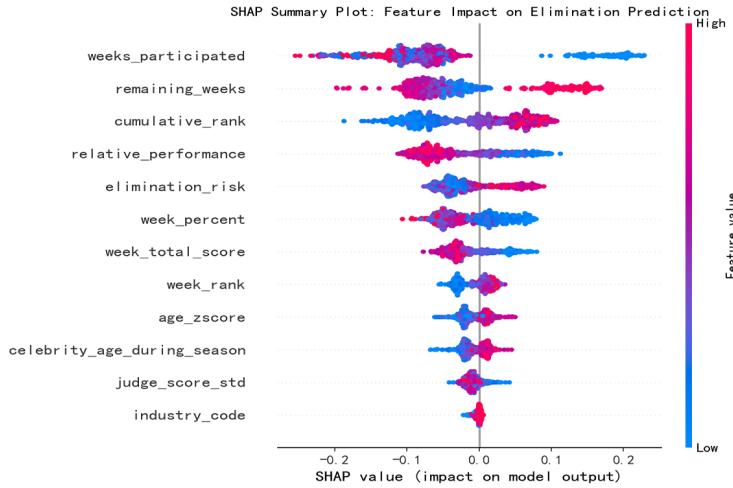


Figure 8: SHAP Summary Plot(Random Forest Model)

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 $\rho = \mathbf{0.8029}$, $p<0.001$, as showed in **Figure 9**), with an **88.76%** agreement rate in prediction direction. This indicates that despite employing fundamentally different methodologies TAN using generative probabilistic modeling and RF using discriminative ensemble learning both 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.

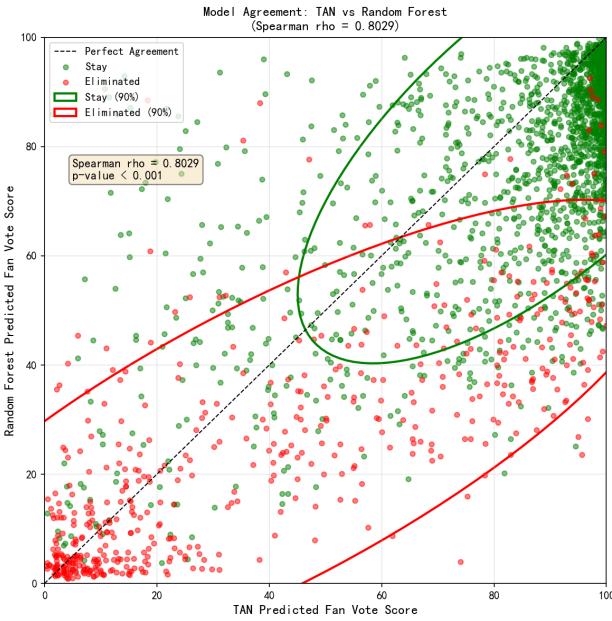


Figure 9: Model Agreement: Random Forest vs TAN

3.4 Validity Test and Robustness Analysis

Stratified 10-fold cross-validation was employed to assess the model's generalization capability. Stratified sampling ensured that the proportion of positive and negative samples in each fold matched the original dataset, thereby mitigating evaluation bias caused by class imbalance.

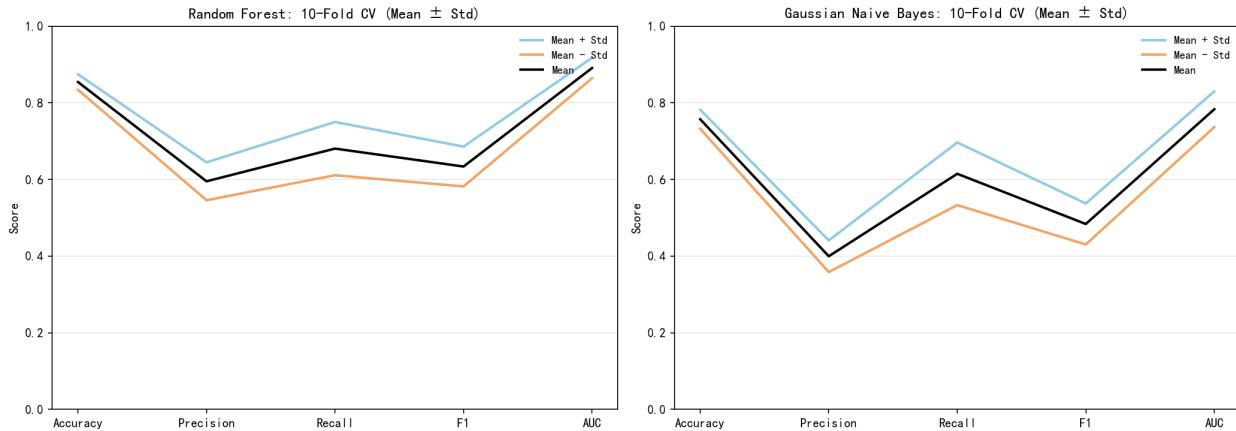


Figure 10: Cross-validation Boxplot

As shown in **Figure 10**, 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.

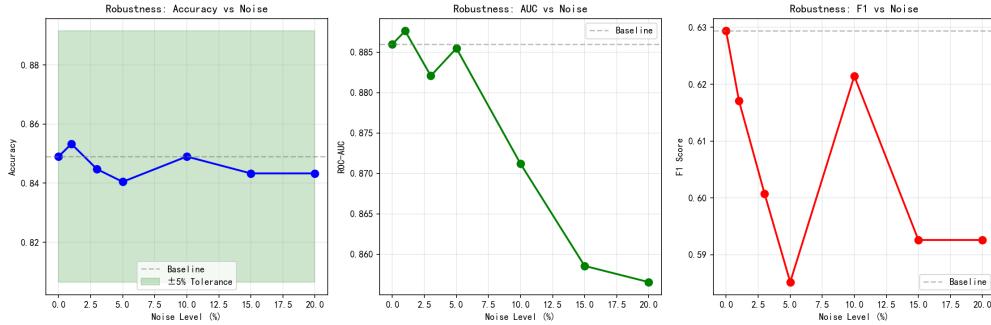


Figure 11: Robustness Analysis: Noise Injection

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.

4 Voting Combinations Comparison via Variance Decomposition Model

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.

$$C_{Rank,i} = R_{judge,i} + R_{fan,i} \quad (8)$$

- 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.

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

4.1 Counterfactual Simulation Modeling

Counterfactual Simulation Modeling Apply both methods to all 34 seasons to construct a controlled 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 12**, 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 12: Counterfactual Simulation Modeling

4.2 Bias Sensitivity Analysis

Bias Sensitivity Analysis aims to quantify the extent to which audience votes influence the final outcome. Decompose the variance of the composite score into judge components and audience components, and calculate the contribution of the audience component to the total variance (audience weighting formula):

$$W_{fan} = \frac{Var(Fan_Component)}{Var(Combined_Score)} \quad (10)$$

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

$$\text{Marginal_Effect} = \frac{\partial P(Y = 1)}{\partial V_i} \quad (11)$$

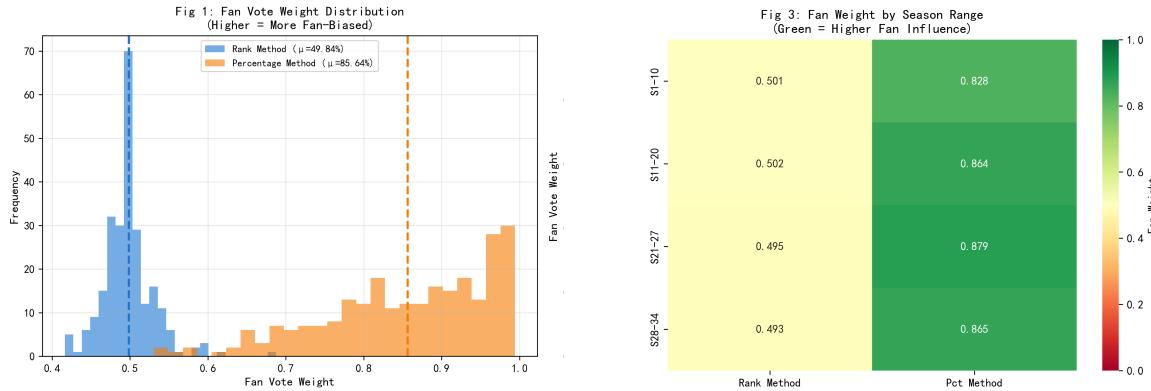


Figure 13: Bias Sensitivity Analysis

As shown in the **Figure 13**, 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 bases such 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.

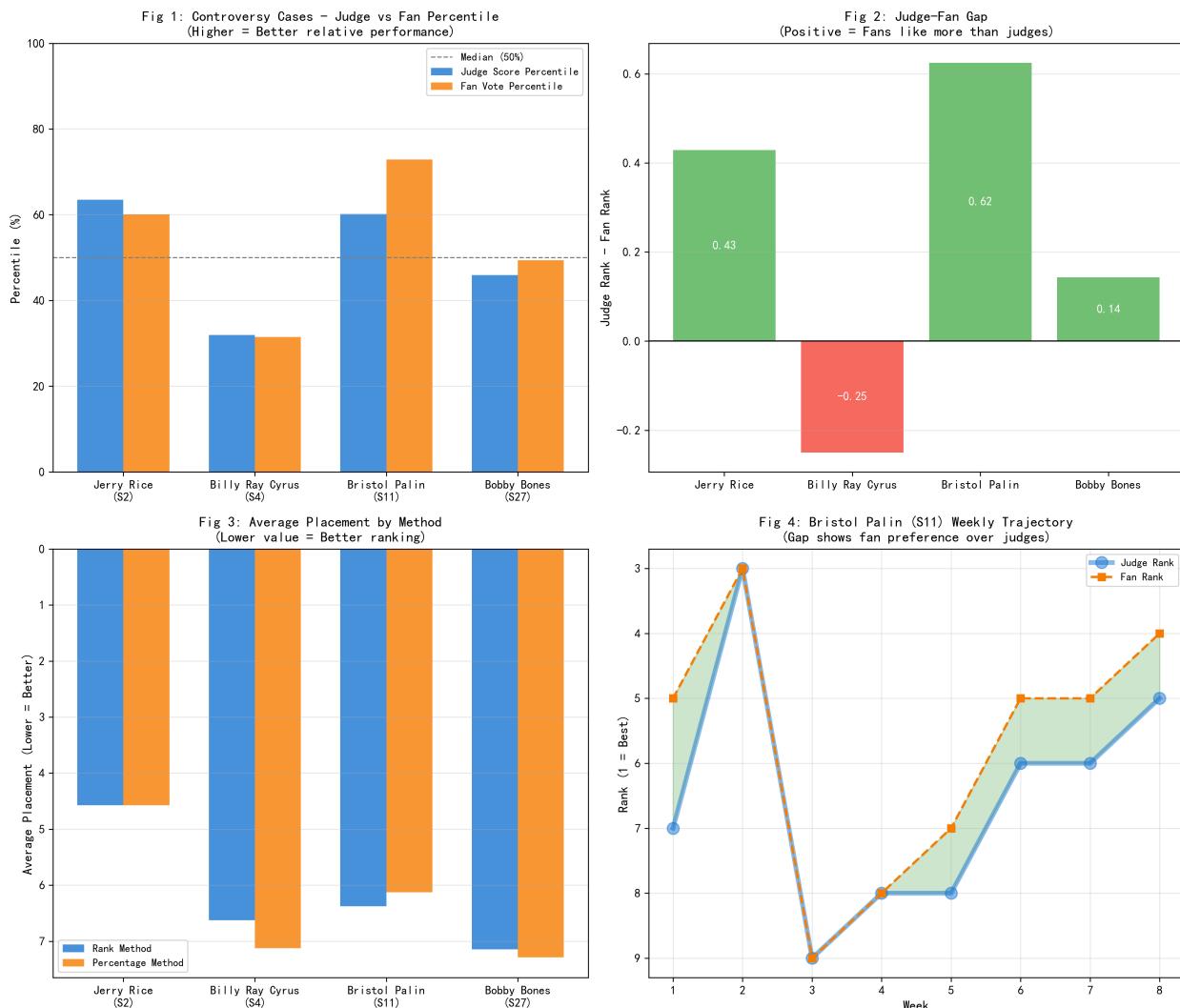


Figure 14: Analysis of Controversial Cases

As shown in the **Figure 14**, 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 15**, thereby **quantifying** the impact of the judges' save mechanism.

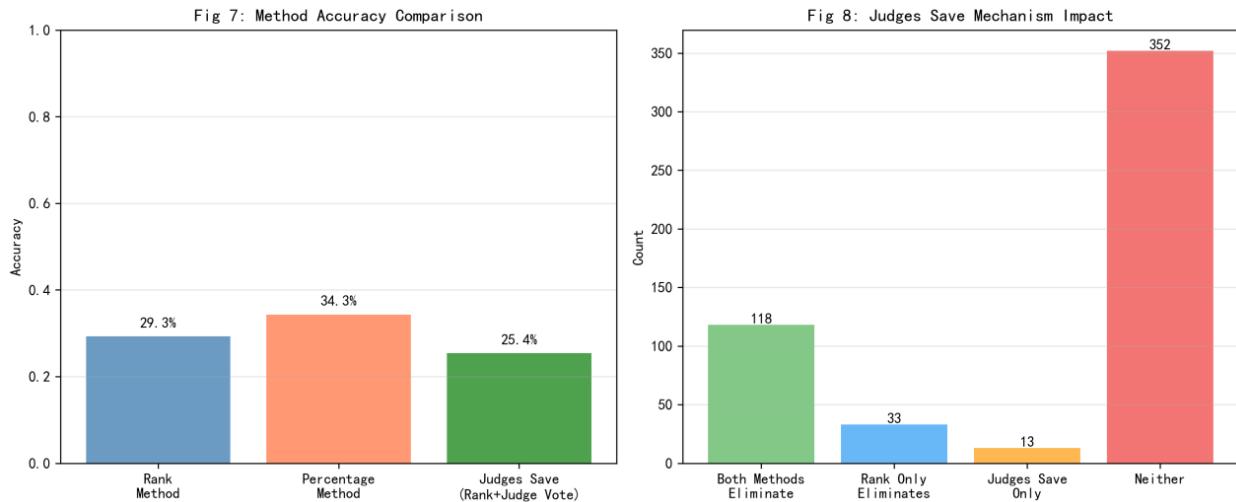


Figure 15: Analysis of the Impact of Judges' Save Mechanism

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.

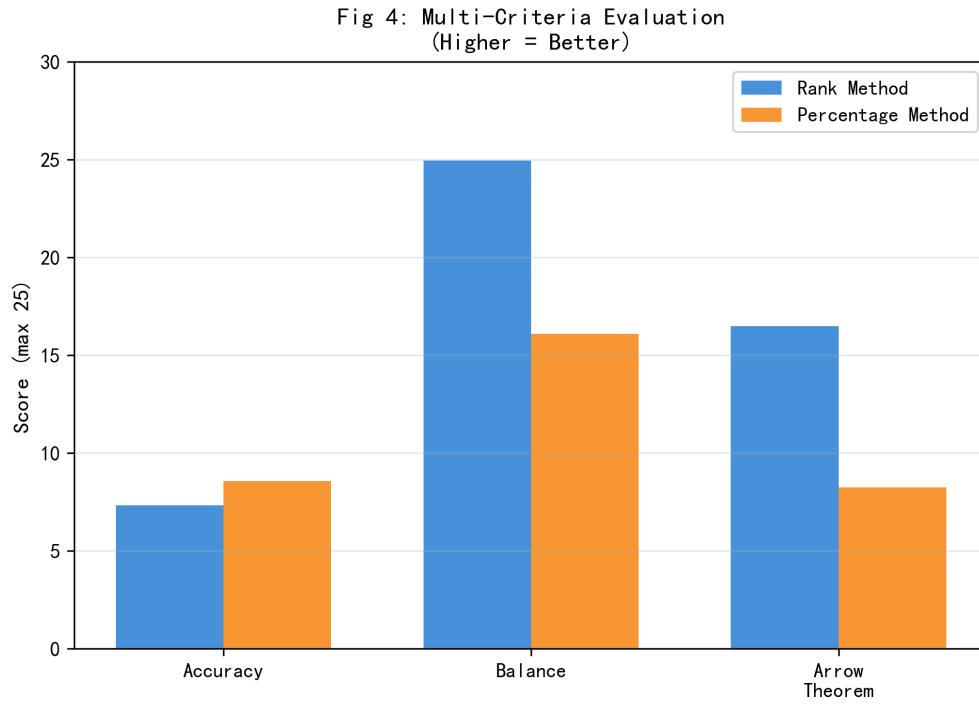


Figure 16: Multi-Criteria Analysis of two Methods

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.

4.5 Multi-Model Validation and Robustness Analysis

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 are listed in the **Table 3** below.

Table 3: 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

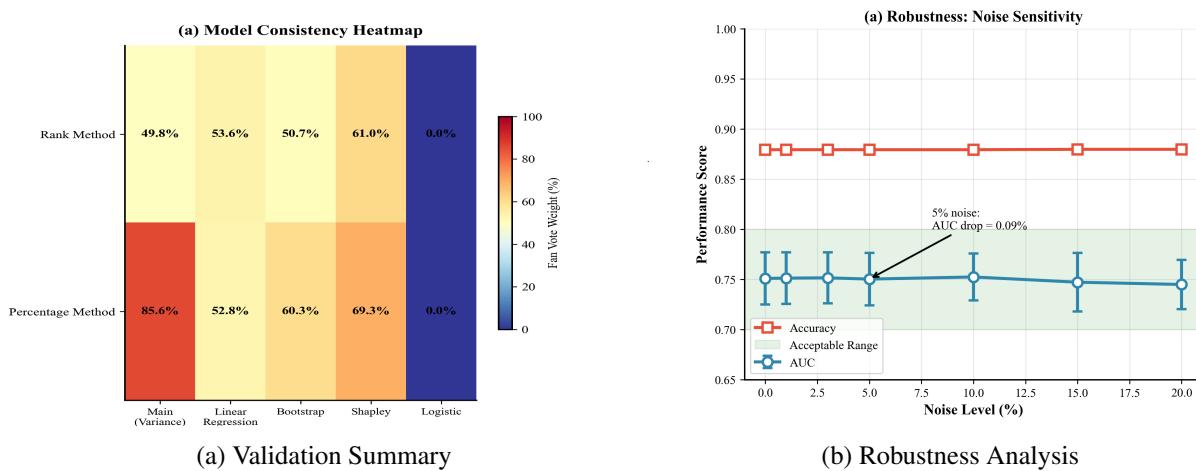


Figure 17: Multi-Model Validation and Robustness Analysis

As shown in the **Figure 17a**, 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 17b**, 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.

5 Quantifying Feature Contributions via Mixed-Effects Model

Problem 3 requires us to analyze the characteristics of dancers and contestants, construct corresponding models, and quantify the contribution of dancer characteristics and contestant characteristics to the contestant's final score and judges' scores.

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. Its basic form is:

$$y_{ij} = X_{ij} \cdot \beta + Z_{ij} \cdot b_j + \epsilon_{ij} \quad (12)$$

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:

$$b_j = \bar{y}_j - \bar{y} \quad (13)$$

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

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

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 4**. 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 4: 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*):

$$ICC = \frac{\sigma_b^2}{\sigma_b^2 + \sigma_{\varepsilon}^2} \quad (15)$$

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

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 18**.

Figure 7: Enhanced Model Effects Analysis

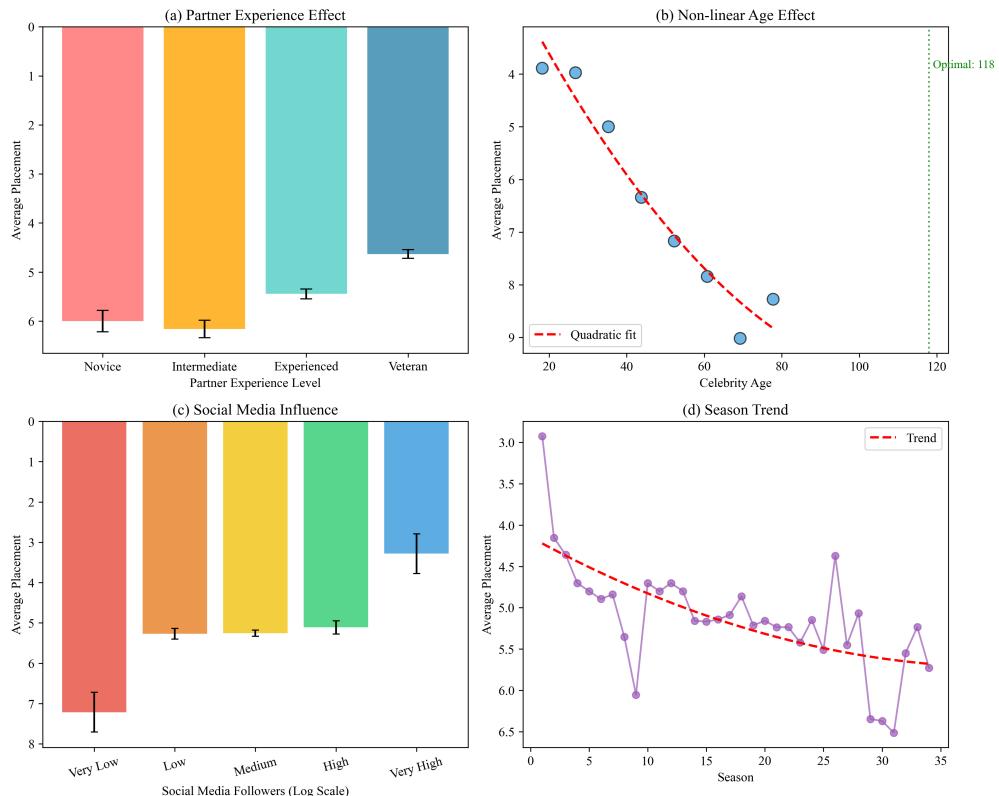


Figure 18: Enhanced Model Effects Analysis

5.3 An Investigation into the Influence of Judges and Audience Voting

A directed acyclic graph (DAG) is a visualization tool used to represent causal relationships between variables. DAG explicitly distinguish between direct effects, indirect effects, and confounding effects. Based on mixed-effects model coefficients and domain knowledge, they identify 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 19** 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 20**, 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.

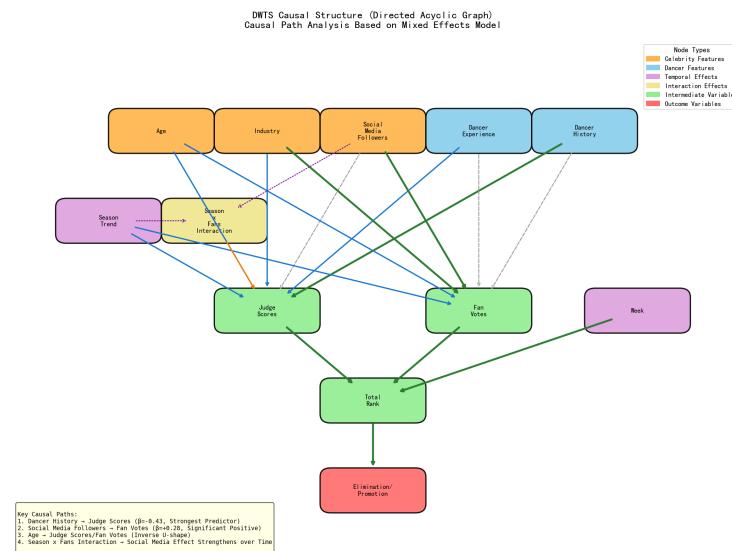


Figure 19: Causal DAG

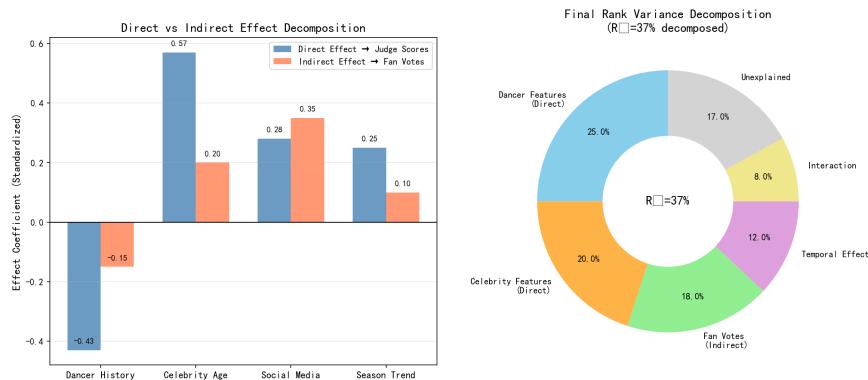


Figure 20: Effect Decomposition

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. As shown in **Figure 21**, 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.

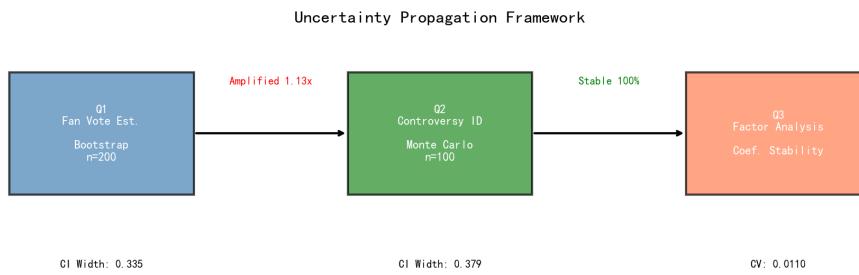


Figure 21: Uncertainty Propagation Framework

Figure 24 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.

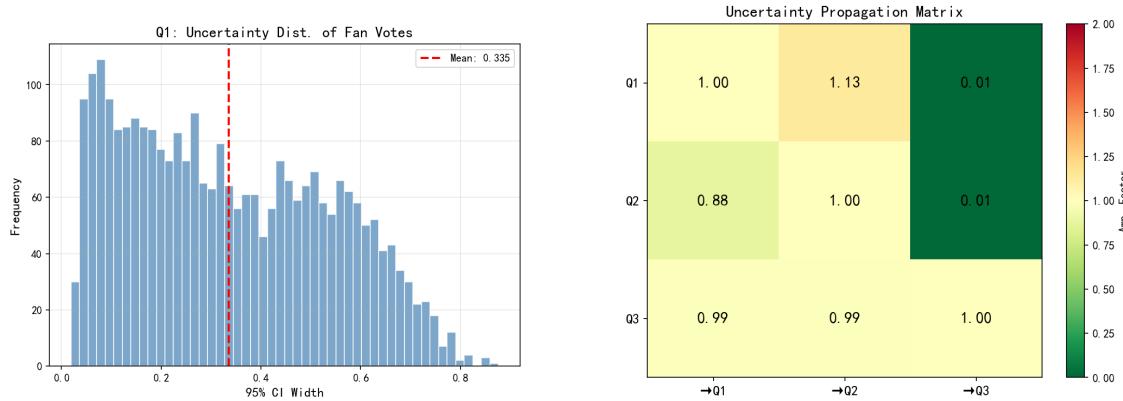


Figure 22: Uncertainty Propagation Analysis Results

7 Design of a Score Fusion Mechanism Based on the KMDR Model

7.1 Establishment and Evaluation of the Score Fusion Mechanism

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.

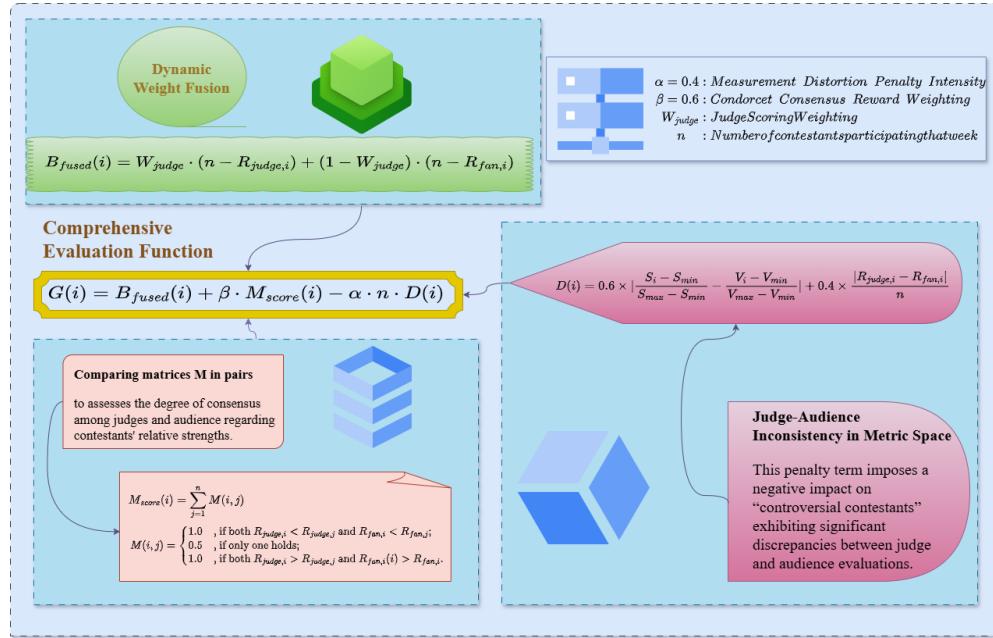
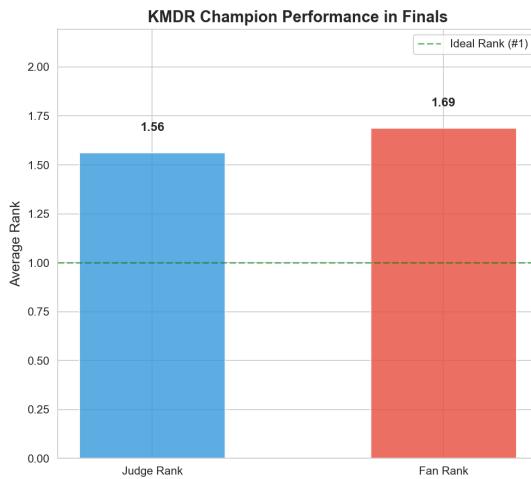


Figure 23: KMDR Model Principle Workflow

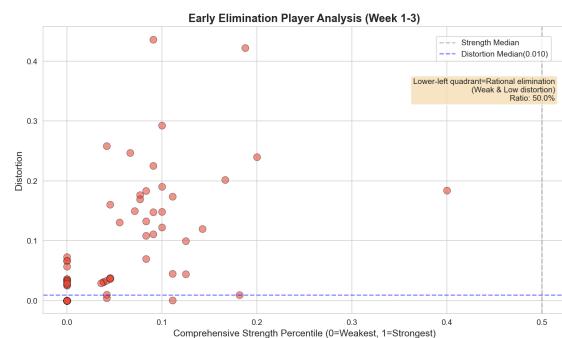
As shown in **Figure 23**, our goal in establishing the model is to design a composite scoring function $G = f(S, V, t)$ that satisfies the fairness, dynamic adaptability, accuracy, and interpretability. Below is the solution and performance evaluation of the scoring fusion mechanism.

- Superiority of Simulation Results

Based on simulations using KMDR, the results are shown in the **Figure 24a**. Across the simulated 32 seasons, we analyzed the quality of these champions from two dimensions: overall strength and controversy. 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 selections a 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 24: KMDR Simulation Analysis Results

- Highly Rational Elimination Decision

The rationality of early elimination decisions serves as a key metric for evaluating ranking systems. As shown in the **Figure 24b**, 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.

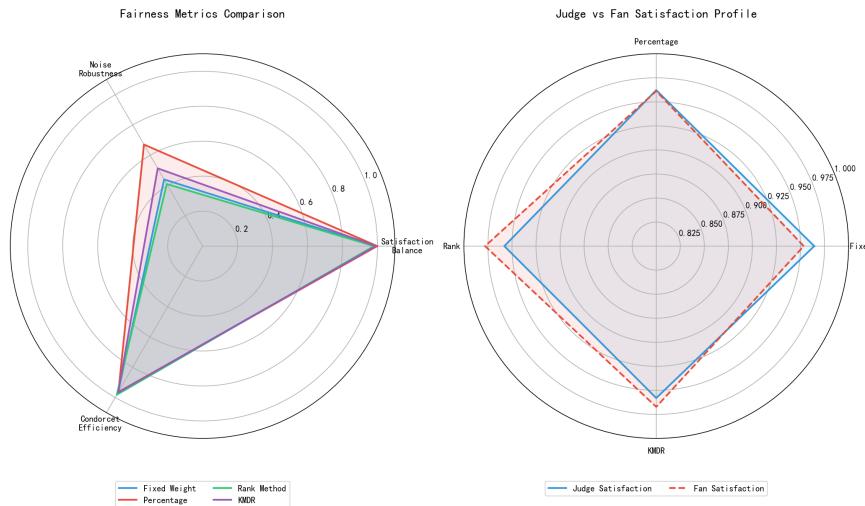


Figure 25: KMDR Fairness Metrics Analysis Results

As shown in the **Figure 25**, 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.

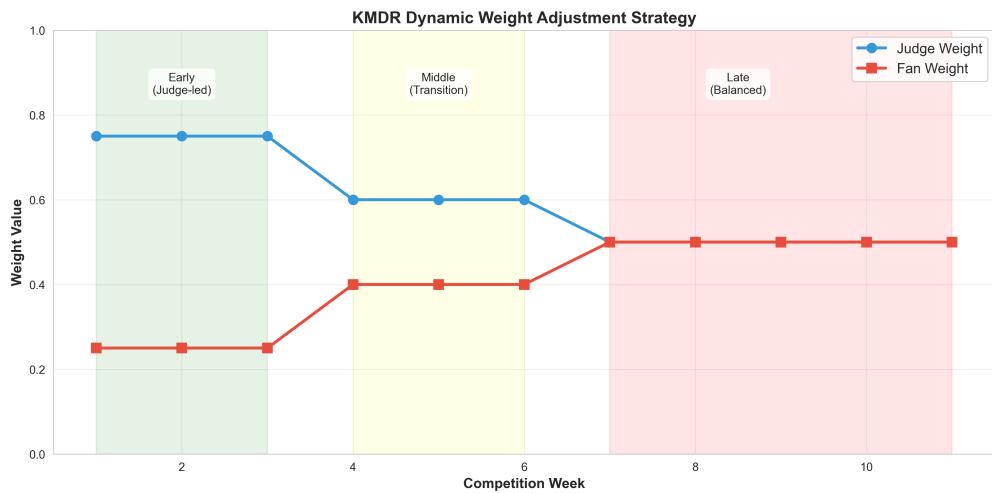


Figure 26: KMDR Dynamic Weight Adjustment Adaptation

- Dynamic Weight Adjustment Adaptation

One of KMDR's core innovations is dynamically adjusting the judge-audience weighting based on the competition stage. **Figure 26** 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 dynamic weighting adjustment strategy 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.

7.2 Sensitivity Analysis

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 ($\alpha=0.1, \beta=0.2$), 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.

8 Sensitivity and Robustness Analysis

9 Strengths and Weaknesses

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] OpenAI. *ChatGPT: Optimizing Language Models for Dialogue*. 2024. URL: <https://openai.com/chatgpt> (visited on 01/08/2024).

A Letter to Community Members

Dear Esteemed Members of Tokyo Tower Community,

We are researchers dedicated to the study of architectural preservation. In order to realize the preservation of Tokyo Tower at a lower cost, please allow me to introduce our proposal on behalf of our team.

We have recently concluded a comprehensive analysis aimed at ensuring the future sustainability and preservation of our cherished Tokyo Tower. Our findings offer a strategic blueprint that prioritizes not only the physical well-being of this iconic structure but also its cultural and historical essence.

Recommendation Plan:

Our proposed plan encompasses innovative preservation techniques, structural enhancements, and community-centric initiatives designed to safeguard and celebrate the tower's legacy. We aim to implement cutting-edge solutions that address both current vulnerabilities and future challenges.

Implementation Timeline:

This ambitious project is scheduled over a five-year period, beginning with immediate preliminary assessments and followed by successive phases of structural upgrades and community engagement activities. This phased approach ensures meticulous attention to detail and the successful realization of our objectives.

Budget Overview:

The financial blueprint for this venture is estimated at \$320,000 per annum. This budget encompasses costs associated with structural reinforcements, preservation technology, and community enrichment programs. It represents a balanced investment in tower's longevity and its significance to our community.

We stand at a pivotal moment in the history of Tokyo Tower, poised to embark on a journey that not only secures its future but also reinvigorates its role within our community. We invite you to support this noble cause, ensuring that Tokyo Tower remains a symbol of resilience, heritage, and communal unity for countless years to come.

Sincerely yours,
Team #2400860

Appendices

Input Python source:

```
import numpy as np

def fft(x):
    n = len(x)
    if n == 1:
        return x
    even = fft(x[0::2])
    odd = fft(x[1::2])
    T = [np.exp(-2j * np.pi * k / n) * odd[k] for k in range(n // 2)]
    return [even[k] + T[k] for k in range(n // 2)] + [even[k] - T[k] for k in
                                                       range(n // 2)]
```

Input Matlab source:

```
function [t,seat,aisle]=OI6Sim(n,target,seated)
pab=rand(1,n);
for i=1:n
    if pab(i)<0.4
        aisleTime(i)=0;
    else
        aisleTime(i)=trirnd(3.2,7.1,38.7);
    end
end
```

Report on Use of AI

1. OpenAI ChatGPT [4] (Jan 25, 2024 version, ChatGPT-4)

Query1: Give me some brainstorming ideas for innovative data visualization techniques.

Output: Here are some innovative data visualization technique ideas:

(1) Interactive Dashboards: Develop dashboards that allow users to interact with the data, such as filtering, zooming, and selecting specific data points to see detailed information.

(2) Animated Visualizations: Use animations to show changes in data over time, helping to illustrate trends, cycles, or progressions.

(3) Augmented Reality (AR) Visualizations: Implement AR to overlay data on real-world objects or environments, enhancing understanding of spatial and contextual relationships.

(4) 3D Visualizations and Models: Employ 3D models to represent complex data structures, allowing exploration from different angles and perspectives.

(5) Narrative Visualizations: Combine data visualization with storytelling, guiding viewers through a narrative that explains the significance of the data in a compelling way.

(6) Real-time Data Streaming: Visualize data in real-time, updating visuals as new data comes in, ideal for monitoring and tracking systems.

(7) Data Sonification: Convert data points into audio signals, making it possible to “listen” to the data for patterns, trends, or anomalies.

These ideas can help present data in more engaging, insightful, and accessible ways.

2. OpenAI ChatGPT (Jan 25, 2024 version, ChatGPT-4)

Query1: What is Engineering News-Record? Give me a brief introduction.

Output: Engineering News-Record (ENR) is a prominent publication within the engineering and construction industry, providing news, analysis, commentary, and data vital for professionals in the field. The publication's roots can be traced back to two separate publications from the late 19th century, eventually merging into what we know today as ENR in 1917. It has evolved to offer a wide range of content including top design firms, construction companies, projects in the U.S. and around the world, and industry trends affecting business management, design, construction methods, technology, and more. ENR is recognized for its comprehensive rankings of the largest construction and engineering firms, based on gross revenues, and for its ENR Award of Excellence, which honors individuals who have made significant contributions to the industry. Today, ENR is owned by BNP Media and continues to serve as an essential resource for contractors, project owners, engineers, architects, and other industry stakeholders with its print and digital magazine, website, newsletters, webinars, and events.

3. OpenAI ChatGPT (Jan 25, 2024 version, ChatGPT-4)

Query1: What's the ENR for?

Output: The Engineering News-Record (ENR) serves several vital functions in the engineering and construction industry. It is recognized for:

(1) News and Analysis: ENR provides comprehensive news coverage, analysis, and commentary on various topics relevant to the construction industry, including major projects, industry trends, technological advancements, and issues affecting the sector such as safety, law, and labor.

(2) Rankings and Data: It publishes annual rankings of the largest construction and engineering firms based on gross revenues, including lists like the Top 400 U.S. General Contractors, Top 500 U.S. Design Firms, and Top 600 Specialty Contractors, among others. These rankings are widely respected and used as benchmarks within the industry.

(3) Awards and Recognition: ENR recognizes industry excellence through various awards, including the ENR Award of Excellence, which honors individuals and projects that have made significant contributions to the industry.

(4) Industry Insights: The publication offers insights on business management, project delivery, technology, workforce development, and other critical areas impacting the engineering and construction fields.

ENR serves as a key resource for professionals across the construction industry, including contractors, project owners, engineers, architects, and suppliers, connecting diverse sectors with coverage that addresses common challenges and opportunities.