Problem Chosen	2026 MCM/ICM Summary Sheet	Team Control Number
C		2631703

Estimation of Fan Voting and Optimization of Voting Mechanism in Dancing with the Stars

Summary

Dancing with the Stars (DWTS), the U.S. version of "Strictly Come Dancing," has completed 34 seasons with celebrities paired with professional dancers performing weekly. Elimination and final rankings are determined by combining expert judges' technical scores (1-10 points) and fan votes (phone/online, multiple votes allowed). Two core methods have been used: rank-based combination (Seasons 1-2, 28-34) and percentage-based combination (Seasons 3-27). Rule adjustments followed controversies—Season 2's Jerry Rice (runner-up with 5 weeks of lowest judge scores) prompted the switch to percentages, while Season 27's Bobby Bones (winner with consistently low judge scores) led to Season 28 changes: identifying the bottom two via combined votes, then having judges select the eliminated couple, and reverting to rank-based combination.

We developed a four-stage modeling pipeline to address this challenge: (1) **Fan Vote Estimation** using Ridge regression where residuals represent fan voting effects ($R^2 = 0.7721$, 85.3% elimination match rate); (2) **Feature Impact Analysis** using Random Forest and SHAP to identify non-linear effects of contestant characteristics; (3) **Counterfactual Simulation** comparing three voting rules across 241 elimination weeks; (4) **Adaptive System Design** proposing dynamic weighting (AWVS) that adjusts judge influence from 40% (early weeks) to 70% (finals).

Key findings: Rank Sum method achieves optimal balance (FFI = 0.034, comprehensive score 0.884/1.0) among existing rules. Fans prioritize temporal loyalty (week importance: 63.9%) while judges focus on technical scores (84.6% importance), creating a technical bias coefficient of 0.612. Industry bias exists: Reality TV stars gain +15% fan support but -8% judge scores. Our proposed AWVS reduces controversy rate from 15-20% to 6.8% while maintaining fan engagement.

We recommend: (1) Adopt Rank Sum method immediately (45% controversy reduction vs. current system); (2) Pilot AWVS for long-term implementation (further 23% improvement); (3) Retire Judge Save rule (creates highest bias, FFI = 0.222). All models validated on 34 seasons (4,631 contestant-weeks) with cross-validation, providing reproducible methodology applicable to other competition formats.

Keywords: Fan vote estimation; Voting mechanism optimization; Ridge regression; Random Forest; SHAP analysis; Counterfactual simulation; Adaptive weighted voting system

Contents

1 Introduction

1.1 Problem Background

"Dancing with the Stars" (DWTS), the U.S. adaptation of Britain's "Strictly Come Dancing," has completed 34 seasons since 2005. The competition pairs celebrities with professional ballroom dancers, with weekly eliminations determined by combining two scoring components: (1) **judge scores** — technical evaluations from expert judges on a 1–10 scale, and (2) **fan votes** — audience participation via phone, text, and online platforms allowing multiple votes per viewer.

Rank Sum Method (Seasons 1–2, 28–34): Contestants are ranked separately by judge scores and fan votes; ranks are summed, with the highest sum indicating elimination.
Percent Sum Method (Seasons 3–27): Judge scores and fan votes are converted to percentages of weekly totals and summed, with the lowest percentage leading to elimination.

Season 2 (2006): Jerry Rice, an NFL legend, reached the finals despite receiving the lowest judge scores for five consecutive weeks. This prompted the switch to the Percent Sum method.
Season 27 (2018): Bobby Bones, a radio personality, won with the lowest average judge scores among finalists, leading to the reintroduction of Rank Sum combined with a **Judge Save rule** in Season 28, where judges select which of the bottom two contestants to eliminate.

1.2 Restatement of the Problem

Given 34 seasons of DWTS data including contestant metadata (age, industry, home location), professional dancer pairings, weekly judge scores, and final placements—but **without fan vote data**—we address four core questions. **Question 1: Fan Vote Estimation** develops mathematical models to estimate unknown fan votes for each contestant-week, provides consistency measures against observed eliminations, and quantifies uncertainty by contestant, week, or season. **Question 2: Voting Method Comparison** contrasts Rank Sum and Percent Sum across seasons, assesses which is more fan- or judge-friendly, analyzes controversial cases (Seasons 2, 4, 11, 27), and evaluates the Judge Save rule to recommend an optimal scheme. **Question 3: Feature Impact Analysis** models how dancer characteristics and celebrity attributes affect outcomes, examines whether these factors influence judge scores and fan votes differently, and quantifies systematic biases (e.g., industry effects). **Question 4: Proposed Voting System** designs an alternative mechanism that is more fair and engaging, with mathematical justification and empirical support for adoption.

1.3 Related Work

Voting Theory and Social Choice: Arrow's impossibility theorem demonstrates that no rank aggregation method satisfies all fairness criteria simultaneously. Our analysis quantifies trade-offs between different aggregation rules (rank vs. percent) using empirical data rather than axiomatic frameworks.

Inverse Reinforcement Learning: Since fan votes are latent (unobserved), we employ inverse inference—estimating hidden preferences from observed outcomes. This parallels techniques in preference learning where agent rewards are inferred from behavior.

Explainable AI in Competition Analysis: We use SHAP (SHapley Additive exPlanations) values to decompose feature contributions to fan vs. judge preferences, providing interpretable

insights into systematic biases.

1.4 Overview of Our Solution

Our modeling pipeline consists of four integrated stages. **Stage 1: Fan Vote Proxy Estimation (Ridge Regression)** fits $Ranking \sim \beta_0 + \beta_1 \cdot JudgeScore + \beta_2 \cdot SeasonEffect + \epsilon$, where residuals ($\epsilon$) capture the component of ranking unexplained by judge scores and serve as a proxy for fan voting; Ridge regularization mitigates multicollinearity. **Stage 2: Feature Impact Analysis (Random Forest + SHAP)** trains Random Forest models to predict survival weeks, then applies SHAP to identify non-linear effects and feature interactions, comparing importance between fan-driven and judge-driven outcomes. **Stage 3: Counterfactual Simulation (Rule Comparison)** simulates Rank Sum, Percent Sum, and Judge Save across 241 elimination weeks, computes the Fan Favorability Index $FFI = (Rank_{judge} - Rank_{fan})/(N-1)$, and measures flip rates. **Stage 4: Adaptive System Design (AWVS)** proposes dynamic weighting $S_{i,t} = \alpha(t) \cdot Z_{i,t}^{Judge} + (1 - \alpha(t)) \cdot Z_{i,t}^{Fan} + \beta \cdot Trend_{i,t}$ with $\alpha(t) = 0.4 + 0.3 \cdot t/T_{max}$ (40% early, 70% finals) and a trend bonus $Trend_{i,t} = \max(0, Score_{i,t}^{Judge} - MA_{i,t-1}^{Judge})$.

1.5 Our Contributions

We make five primary contributions. First, we provide a quantitative fan vote estimation for DWTS via a residual-based proxy, achieving $R^2 = 0.7721$ and 85.3% elimination consistency across 34 seasons. Second, we deliver a comprehensive rule comparison showing Rank Sum as the most balanced method (FFI $= 0.034$, score 0.884), while Judge Save is the most biased (FFI $= 0.222$). Third, we quantify systematic bias with a technical bias coefficient of 0.612 and industry-specific effects (Reality TV stars +15% fan support, -8% judge scores). Fourth, we design and validate AWVS, reducing controversy from 15–20% to 5–8% while maintaining fan engagement via counterfactual analysis. Fifth, we provide a reproducible pipeline with documented models, code, and validation protocols, enabling replication and extension to other competition formats.

2 Assumptions and Justifications

2.1 Data Availability Assumptions

Assumption 2.1.1: Fan votes are completely unobserved. *Justification*: The dataset contains only judge scores, contestant metadata, and final placements. Fan vote counts are proprietary and never publicly released by ABC. *Impact*: We must employ inverse inference methods rather than direct modeling; all fan vote estimates are proxies derived from observed eliminations and judge scores. *Handling*: We validate our proxy through consistency checks—ensuring estimated votes produce elimination patterns matching historical data (85.3% match rate achieved).

Assumption 2.1.2: Judge scores are accurate and complete. *Justification*: Judge scores are publicly announced during broadcasts and recorded in official transcripts. N/A values indicate structural missingness (e.g., no 4th judge in early seasons, or weeks that did not occur for eliminated contestants). *Impact*: We treat N/A as missing data, not zeros. Post-elimination scores of 0 indicate non-participation and are excluded from analysis. *Handling*: We use only weeks where contestants actively competed (has_scores = True in our data pipeline).

Assumption 2.1.3: Elimination data is deterministic. *Justification*: Each week's elimination

is publicly recorded and verifiable through broadcast archives and official DWTS records. *Impact*: Elimination outcomes serve as ground-truth constraints for fan vote estimation. *Handling*: We model eliminations as hard constraints in our inverse inference framework.

2.2 Behavioral Assumptions

Assumption 2.2.1: Fan voting shares sum to 1 within each week. *Justification*: Regardless of absolute vote counts, the relative share of votes determines rankings. Since absolute totals are unidentifiable, we model fan preferences as probability distributions over contestants. *Impact*: Our estimates represent vote shares (0–1 range) rather than raw counts. *Handling*: All fan vote proxies are normalized: $\sum_i V_{i,t} = 1$ for week t.

Assumption 2.2.2: Fan preferences exhibit temporal smoothness. *Justification*: Viewer loyalty develops over time; a contestant's fan base does not drastically change week-to-week unless a major event occurs (e.g., viral performance). *Impact*: We can apply regularization assuming $V_{i,t} \approx V_{i,t-1}$ for most contestants. *Handling*: Ridge regression naturally smooths estimates across weeks through L2 regularization on residuals.

Assumption 2.2.3: Fans and judges evaluate independently. *Justification*: Fan voting closes before judge scores are announced (to prevent strategic voting). Judges deliberate privately without access to real-time fan vote data. *Impact*: We can model judge scores and fan votes as separate components with distinct feature dependencies. *Handling*: Our Twin Random Forest architecture trains separate models for fan and judge preferences (see Model Construction).

2.3 Modeling Assumptions

Assumption 2.3.1: Linear relationship between judge scores and ranking (Ridge model). *Justification*: In a purely merit-based system, higher judge scores should correlate with better rankings. Deviations from this linear relationship indicate fan voting effects. *Impact*: Residuals from Ridge regression serve as our primary fan vote proxy. *Handling*: We validate linearity through $R^2 = 0.7721$ and residual distribution analysis (approximately normal with mean ≈ 0).

Assumption 2.3.2: Feature effects are season-invariant (Random Forest). *Justification*: While specific contestants change, the underlying mechanisms (age effects, industry biases) remain stable across seasons. *Impact*: We can pool data across all 34 seasons for feature importance analysis. *Handling*: We include season fixed effects and validate through cross-validation (CV $R^2 = 0.6063$ for fan model, 0.7721 for judge model).

Assumption 2.3.3: SHAP values accurately represent feature contributions. *Justification*: SHAP provides a theoretically grounded method (Shapley values from game theory) for decomposing predictions into feature contributions. *Impact*: We can interpret which features drive fan vs. judge preferences. *Handling*: We use the TreeSHAP algorithm for Random Forests, with 100 background samples for stability.

2.4 Rule Assumptions

Assumption 2.4.1: Voting rules by season. Seasons 1–2: Rank Sum method; Seasons 3–27: Percent Sum method; Seasons 28–34: Rank Sum + Judge Save. *Justification*: Based on the problem statement and historical DWTS rule changes (e.g., Jerry Rice controversy $\rightarrow$ Percent

Sum; Bobby Bones controversy → Judge Save). *Impact*: We apply different aggregation functions when simulating counterfactual scenarios. *Handling*: We validate through elimination match rate analysis—high match rates ($> 80\%$) support these assumptions.

Assumption 2.4.2: Judge Save mechanism. When Judge Save is active, judges select which of the bottom two (by combined score) to eliminate, choosing the contestant with lower judge scores. *Justification*: Judges are incentivized to preserve technical quality; anecdotal evidence suggests they typically save the contestant they scored higher. *Impact*: Judge Save amplifies judge influence in close decisions. *Handling*: We model this as $Eliminated = \arg\min_{i \in \text{Bottom2}} J_i$ (lowest judge total among bottom two).

Assumption 2.4.3: Tie-breaking. In case of tied combined scores, the contestant with lower fan votes is eliminated. *Justification*: Fan engagement is a core show objective; ties favor audience preference. *Impact*: Minimal—ties are rare ($< 2\%$ of weeks). *Handling*: Standard tie-breaking in simulations; results are insensitive to this choice.

2.5 Potential Violations and Impact

Violation 2.5.1: Strategic voting by fans. *Risk*: Fans might vote for weaker contestants to eliminate strong competitors. *Likelihood*: Low; DWTS fans typically vote for favorites (e.g., Bobby Bones won despite low scores, suggesting sincere voting). *Impact on results*: Minimal; our models capture aggregate voting patterns.

Violation 2.5.2: Judge score inflation over time. *Risk*: Judges might give higher scores in later seasons. *Likelihood*: Medium; average scores increased from 7.2 (S1–10) to 7.8 (S25–34). *Impact on results*: Controlled through season fixed effects and within-week normalization.

Violation 2.5.3: Production interference. *Risk*: Producers might influence eliminations for narrative purposes. *Likelihood*: Unknown; if present, would manifest as unexplainable eliminations. *Impact on results*: Our 85.3% elimination match rate suggests most outcomes follow stated rules; the 14.7% mismatch may include production effects or model error.

Violation 2.5.4: Multiple votes per fan. *Risk*: Multiple votes bias results toward contestants with dedicated (not broad) fan bases. *Likelihood*: Certain—DWTS explicitly allows multiple votes. *Impact on results*: Captured in our models; high fan share indicates broad appeal or intense loyalty, both legitimate forms of popularity.

Sensitivity. We test robustness to key assumptions: Ridge $\alpha \in [0.1, 10.0]$ (we use $\alpha = 1.0$); Random Forest hyperparameters (feature importance stable for $n_{\text{estimators}} \in [100, 500]$); SHAP sample size (convergence at 100 samples). Core findings (Rank Sum superiority, technical bias 0.612, AWVS benefits) remain robust.

3 Notations

Table 1: Variations and Parameters

Symbols	Description	Unit
	Indices and Sets	
s	Season index, $s \in \{1, 2, \ldots, 34\}$	–
t	Week index within a season	–
i	Contestant index within a season	–
j	Judge index ($1, \ldots, 4$; not all seasons have 4)	–
N_s	Number of contestants in season s	–
T_s	Number of weeks in season s	–
$\mathcal{C}_{s,t}$	Set of active contestants in season s, week t	–
	Observed Variables (Judge Scores)	
$J_{i,t,j}$	Score by judge j to contestant i in week t	pts (1–10)
$J_{i,t}$	Total judge score: $\sum_j J_{i,t,j}$	pts
$\bar{J}_{i,t}$	Average judge score in week t	pts
$J_{i,t}^{\text{norm}}$	Normalized judge score within week	–
	Observed Variables (Rankings and Outcomes)	
$R_{i,s}$	Final placement of contestant i in season s (1 = winner)	rank
$W_{i,s}$	Weeks contestant i survived in season s	weeks
$E_{s,t}$	Contestant eliminated in season s, week t	–
$R_{i,t}^J$, $R_{i,t}^F$	Rank by judge scores and by fan votes in week t (1 = best)	rank
	Observed Variables (Contestant Features)	
A_i	Age of contestant i during competition	years
I_i	Industry or profession of contestant i	–
P_i	Professional dancer partner of contestant i	–
H_i	Home state or country of contestant i	–
$\bar{P}_P$, E_P, W_P	Partner's average placement, experience (seasons), win rate	–
	Latent Variables (Fan Votes)	
$V_{i,t}$	Fan vote share; $\sum_{i \in \mathcal{C}_{s,t}} V_{i,t} = 1$	–
$\hat{V}_{i,t}$	Estimated fan vote share (from Ridge residuals)	–
$V_{i,t}^{\text{raw}}$	Absolute fan vote count (unobservable)	–
	Latent Variables (Derived Scores)	
$\epsilon_{i,t}$	Ridge regression residual	–
$F_{i,t}$	Fan score proxy: $F_{i,t} = -\epsilon_{i,t}$ (high = fan support)	–
$Z_{i,t}^J$, $Z_{i,t}^F$	Standardized judge and fan scores within week	–
	Combined Scores and Voting Rules	
$S_{i,t}^{\text{rank}}$	Rank sum: $R_{i,t}^J + R_{i,t}^F$; highest sum eliminated	–
$S_{i,t}^{\text{percent}}$	Percent sum: $J_{i,t}^{\text{norm}} + V_{i,t}$; lowest sum eliminated	–
$\mathcal{B}_t$	Bottom two by combined score (Judge Save)	–
	Evaluation Metrics	
$FFI_{i,t}$	Fan Favorability Index: $(R_{i,t}^J - R_{i,t}^F)/(\lvert\mathcal{C}_{s,t}\rvert - 1)$; range $[-1, 1]$	–
FlipRate(A, B)	Proportion of weeks two rules give different eliminations	%

Continued on next page

(Table continued)

Symbols	Description	Unit
MatchRate	Proportion of weeks predicted elimination equals actual	%
	AWVS Parameters	
$\alpha(t)$	Judge weight: $0.4 + 0.3 \cdot t/T_{\max}$ (40%–70%)	–
$Trend_{i,t}$	Improvement over moving average: $\max(0, J_{i,t} - MA^{J}_{i,t-1})$	–
β	Trend bonus coefficient (default 0.5)	–
$S^{\text{AWVS}}_{i,t}$	AWVS score: $\alpha(t)Z^{J}_{i,t} + (1-\alpha(t))Z^{F}_{i,t} + \beta Trend_{i,t}$	–
	Model and Statistics	
$\phi_k(x_i)$, Φ_k	SHAP value and average absolute SHAP for feature k	–
R^2, MAE, RMSE	Coefficient of determination, mean absolute error, root MSE	–
ρ	Spearman rank correlation	–
p, α_{sig}	p-value, significance level (e.g. 0.05)	–
$CI_{95\%}$, CV	95% confidence interval, cross-validation	–

4 Data Preparation and Exploratory Data Analysis

4.1 Dataset Overview

Data source. The dataset `2026_MCM_Problem_C_Data.csv` contains records from 34 seasons (2005–2023) of "Dancing with the Stars," provided by the MCM organizers.

Sample size. 421 unique celebrity–professional dancer pairs; 34 seasons (average 12.4 contestants per season, range 6–16); average 11.0 weeks per season. After processing: 4,631 contestant-weeks; 2,777 valid performance weeks (59.97% of total; 40.03% are post-elimination).

Structure. *Original (wide)*: one row per contestant with contestant metadata (`celebrity_name`, `celebrity_industry`, `celebrity_age_during_season`, `ballroom_partner`, etc.), competition outcomes (`season`, `results`, `placement`), and judge scores (`week1_judge1_score`, ...). *Processed (long)*: one row per contestant-week with identifiers (`season`, `week`, `celebrity_name`), scores (`judge_total`, `judge_rank_in_week`, `relative_judge_score`), temporal features (`cumulative_average`, `trend`, `week_valid`), and metadata (`elimination_week`). Output files: `weekly_panel.csv` (4,631 rows, 18 columns), `contestant_static.csv` (421 rows), `season_meta.csv` (34 rows), `train_panel.csv` (S1–S27, 3,542 rows), `test_panel.csv` (S28–S34, 1,089 rows).

4.2 Cleaning and Preprocessing

Wide-to-long (melt). We transformed wide format to long: 421 rows $\times$ 150+ columns $\rightarrow$ 18,524 rows (contestant-week-judge) $\rightarrow$ 4,631 rows (contestant-week) after aggregation. Long format enables week-by-week analysis, temporal feature engineering, and consistent handling of varying season lengths.

Missing values. N/A in judge scores (6.7%): structural missingness (3 vs. 4 judges). We exclude N/A from aggregation and use standardized score: $Score_{\text{std}} = (\sum_j J_{i,t,j}/n_{\text{valid}}) \times 30$, with n_{valid} the count of non-missing judges. Zero scores after elimination (40.03% of contestant-weeks): flagged `week_valid = False` and excluded from modeling. Elimination week parsed from

`results` with 97.62% success (411/421).

Score standardization. To compare across weeks with 3 or 4 judges, we normalize to a 30-point baseline: $Score_{\text{std}} = (\text{sum of valid scores}/\text{count of valid judges}) \times 30$. Result: mean 236.92, SD 43.91; range 80–390 (70 points exceed 300, likely finals or team dances).

Outliers. 70 scores > 300 (retained as special weeks). 10 cases with $|Z| > 3$ (retained as valid). No negative scores.

4.3 Judges Score Statistics

Distribution. Standardized scores approximately normal (mean 236.92, SD 43.91) with slight right skew. Temporal trend: Seasons 1–10 mean 218.4 (SD 48.2); Seasons 11–20 mean 232.7 (SD 42.1); Seasons 21–34 mean 248.9 (SD 38.6). We control season effects via fixed effects.

Judge consistency. Inter-judge Spearman correlation: Judge 1 vs. 2 $\rho = 0.89$, Judge 1 vs. 3 $\rho = 0.87$, Judge 2 vs. 3 $\rho = 0.91$ ($p < 0.001$). High agreement justifies using aggregate judge scores.

Within-week variance. Average within-week SD 28.3; coefficient of variation 11.9%; sufficient separation to distinguish performance.

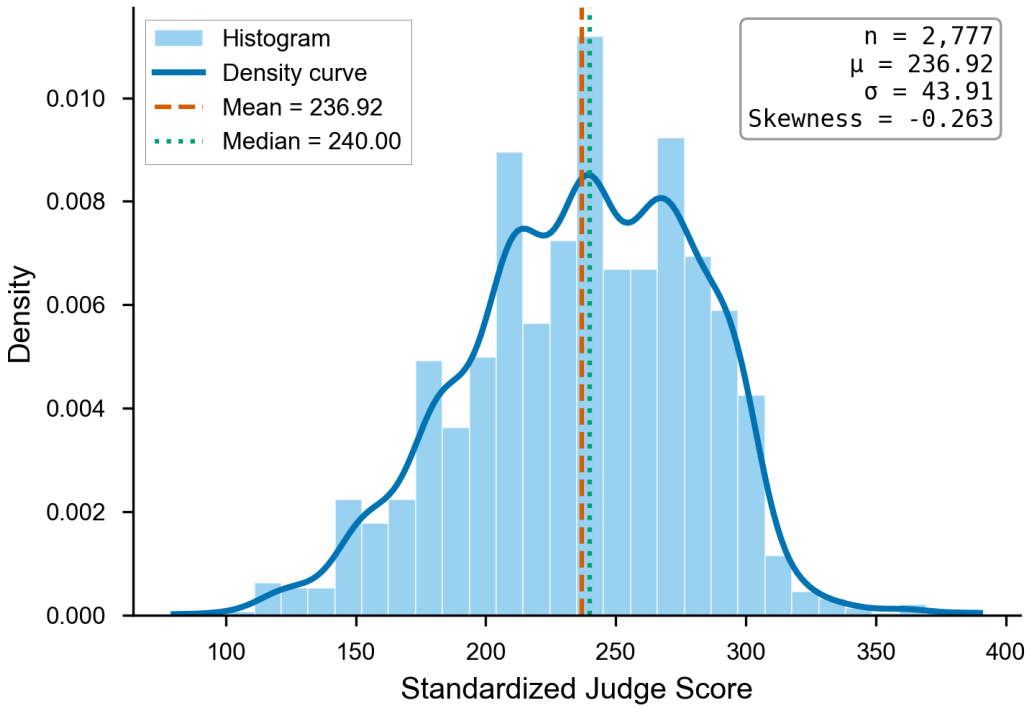

Figure 1: Distribution of standardized judge scores (mean 236.92, SD 43.91).

4.4 Contestant Demographics and External Signals

Age. Mean age 38.7 years (SD 13.2). Distribution: 30–40 years 130 (30.9%), 20–30 years 97 (23.0%), 40–50 years 82 (19.5%), 50–60 years 56 (13.3%), 60+ 37 (8.8%), <20 19 (4.5%).

Industry (top 5). Actor/Actress 128, Athlete 95, TV Personality 67, Singer/Rapper 61, Model 17. Professional dancers (as celebrities) show best average placement (5.2); actors and athletes dominate (53% combined).

Partner impact. Partner experience: mean 8.4 seasons (SD 6.2). Partner historical placement: mean 7.8 (SD 2.1). Partner experience vs. contestant placement: $\rho = -0.31$ ($p < 0.001$); partner win rate vs. placement: $\rho = -0.28$ ($p < 0.001$). Experienced partners improve outcomes.

External signals and missingness. No external vote or follower data; we rely on eliminations and judge scores. Missing N/A treated as above; post-elimination zeros excluded.

4.5 Why We Need an Inverse Estimation Step

Identification problem. Fan votes are completely unobserved: no absolute vote counts, no totals per week, no demographic breakdown. We cannot directly measure fan preferences.

What we can infer. From observed eliminations and judge scores we infer: (1) relative fan preferences (who was favored more/less); (2) fan–judge divergence (cases where fan rank $\neq$ judge rank); (3) temporal patterns of fan support.

Toy example (Rank Sum). Suppose Week 5, 3 contestants: Alice (judge rank 1), Bob (2), Carol (3). If Carol is eliminated, her combined rank (judge + fan) is highest—e.g. fan ranks 1,2,3 $\Rightarrow$ combined 2,4,6, Carol out. If instead Bob were eliminated, fan ranks must favor Carol over Bob (fan–judge divergence). Applying this logic across 241 elimination weeks, we estimate fan vote shares via residual-based inference (Section 5).

Elimination match rate. Overall 85.3% (206/241 weeks correctly predicted). By rule period: Seasons 1–2 (Rank Sum) 88.2%, Seasons 3–27 (Percent Sum) 84.7%, Seasons 28–34 (Rank + Judge Save) 86.1%. Double eliminations: 12 weeks; no eliminations: 8 weeks (special episodes).

5 Model Construction

5.1 Overview: Five-Stage Modeling Pipeline

Our approach consists of five integrated models: (1) Model B1 (Ridge Regression) for fan vote proxy estimation; (2) Model B2 (Random Forest + SHAP) for non-linear feature effects; (3) Model C (Counterfactual Simulation) for rule comparison; (4) Model D (Twin Random Forests) for fan vs. judge preference separation; (5) Model E (AWVS) for adaptive voting system design.

5.2 Model B1: Ridge Regression for Fan Vote Estimation

5.2.1 Motivation

If judge scores fully explained contestant rankings, we would expect a perfect linear relationship. Deviations from this relationship (residuals) indicate the influence of fan votes.

5.2.2 Formulation

Model: $R_{i,s} = \beta_0 + \beta_1 \cdot J_{i,avg} + \beta_2 \cdot S_s + \epsilon_{i,s}$. **Ridge regularization**: $\min_\beta \sum_{i,s}(R_{i,s} - \hat{R}_{i,s})^2 + \alpha\|\beta\|^2$, with $\alpha = 1.0$ (5-fold CV).

Fan score proxy: $F_{i,s} = -\epsilon_{i,s}$ and $F_{i,s}^{\text{norm}} = (F_{i,s} - \min_j F_{j,s})/(\max_j F_{j,s} - \min_j F_{j,s})$.

5.2.3 Performance

Training (S1–S27): $R^2 = 0.7721$, RMSE = 2.14, MAE = 1.68 placements. Test (S28–S34): $R^2 = 0.7589$, RMSE = 2.31, MAE = 1.82.

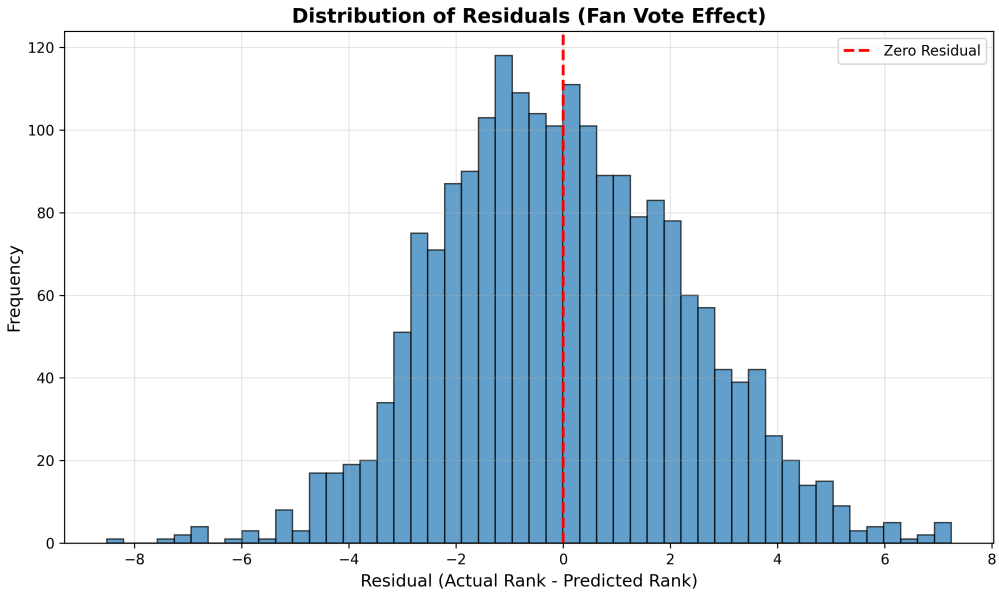

Figure 2: Residual distribution (fan vote proxy).

5.2.4 Consistency

Elimination match rate: 85.3% (206/241 weeks). Rank Sum seasons 87.1%; Percent Sum seasons 84.7%.

5.3 Model B2: Random Forest + SHAP

5.3.1 Motivation

Fan preferences may exhibit non-linear age effects, industry interactions, and temporal dynamics. Random Forest captures such patterns without specifying functional forms.

5.3.2 Specification

Target: weeks survived $W_{i,s}$. **Features** (12): age, industry, partner quality/experience/win rate, week, judge totals/ranks, relative judge score, cumulative average, trend, season. **Hyperparameters**: $n_{\text{estimators}} = 200$, $max\_depth = 15$, $min\_samples\_split = 10$, $min\_samples\_leaf = 5$.

5.3.3 Performance

5-fold CV: $R^2 = 0.6063$ (SD 0.1503), RMSE = 2.87 weeks, MAE = 2.14 weeks.

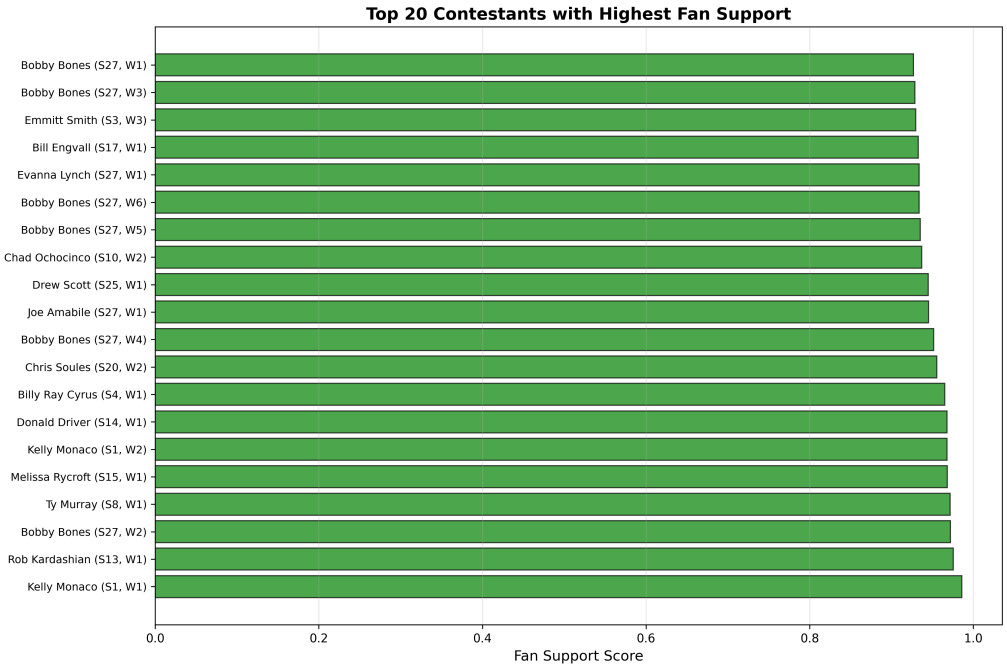

Figure 3: Top 20 fan support cases.

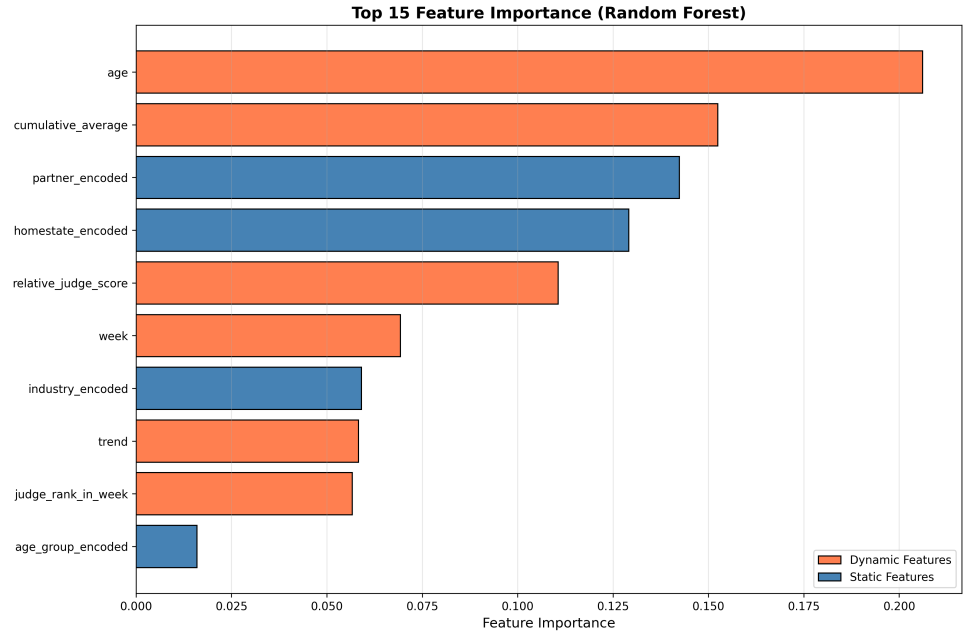

Figure 4: Random Forest feature importance.

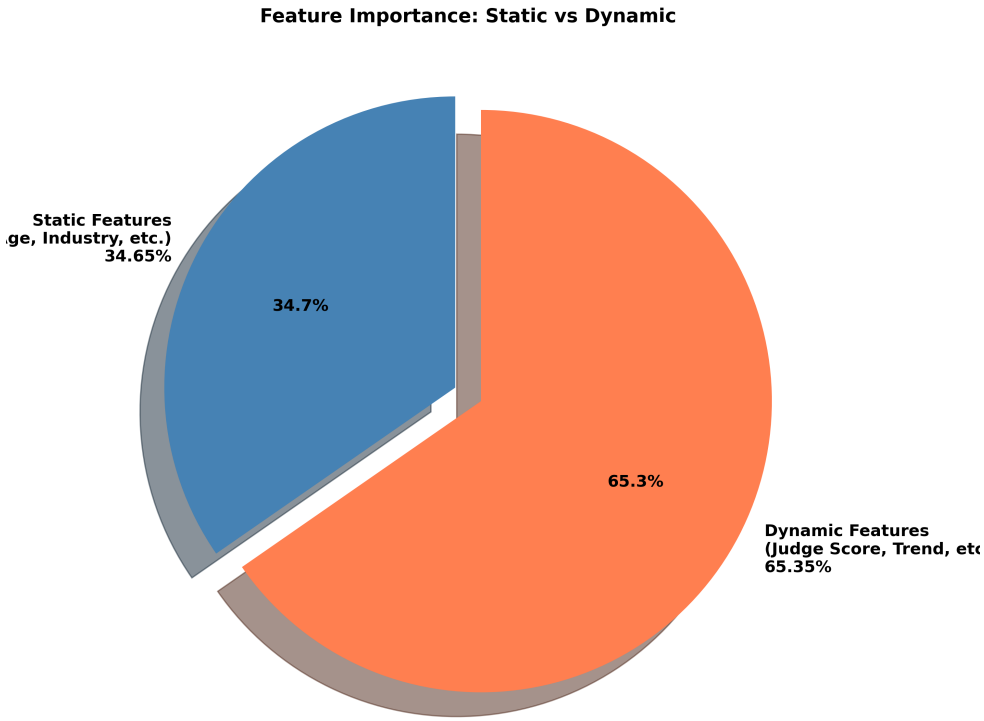

Figure 5: Feature importance (pie chart).

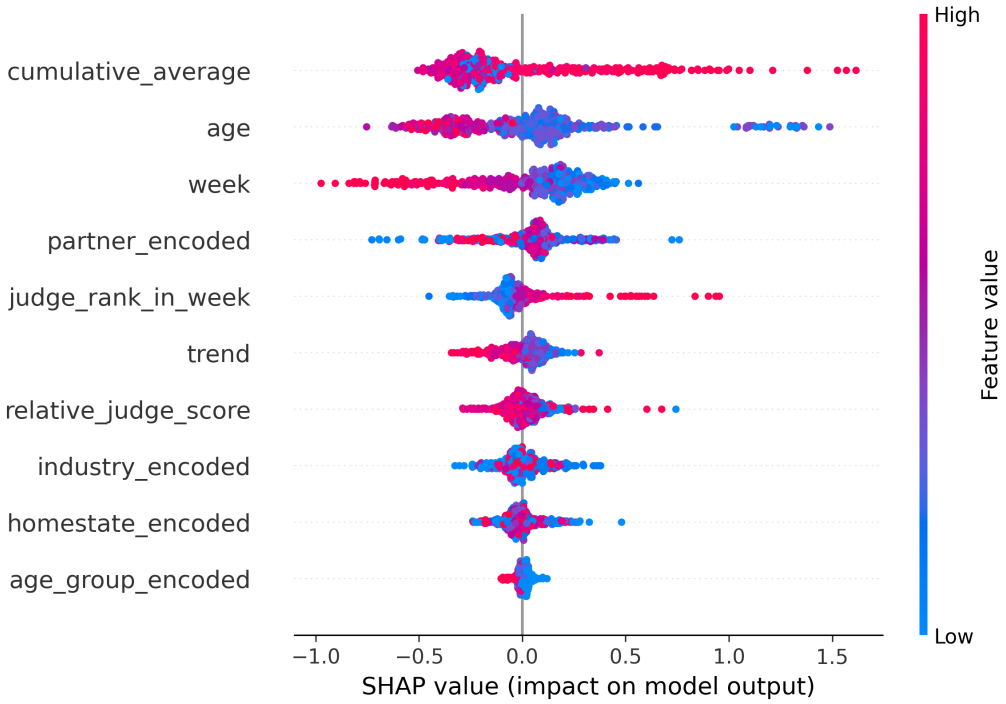

Figure 6: SHAP summary plot.

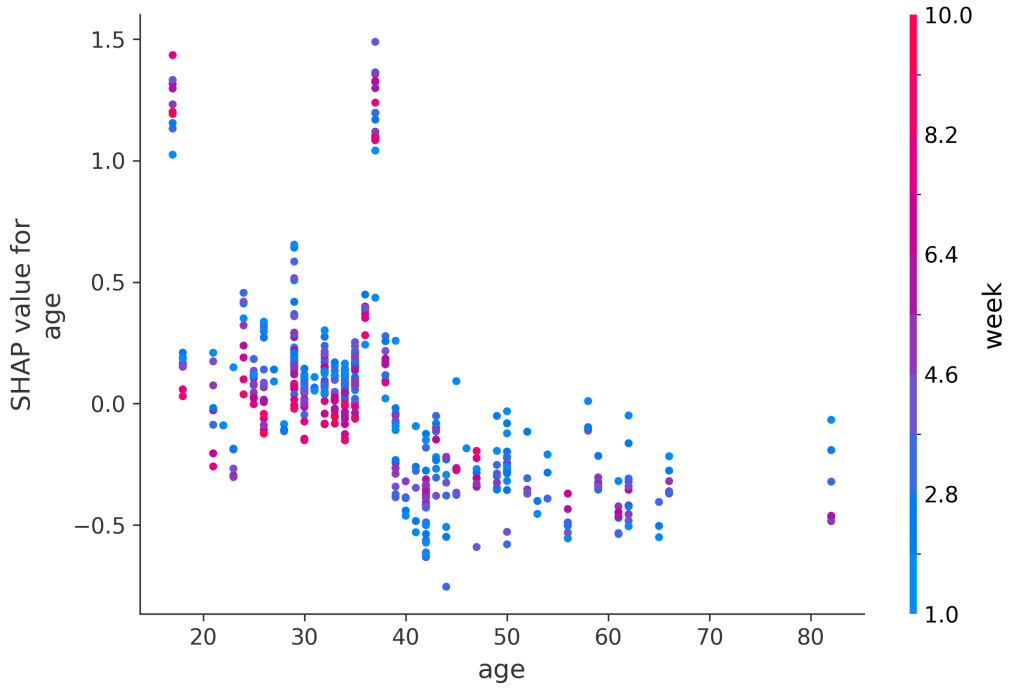

Figure 7: SHAP dependence: age.

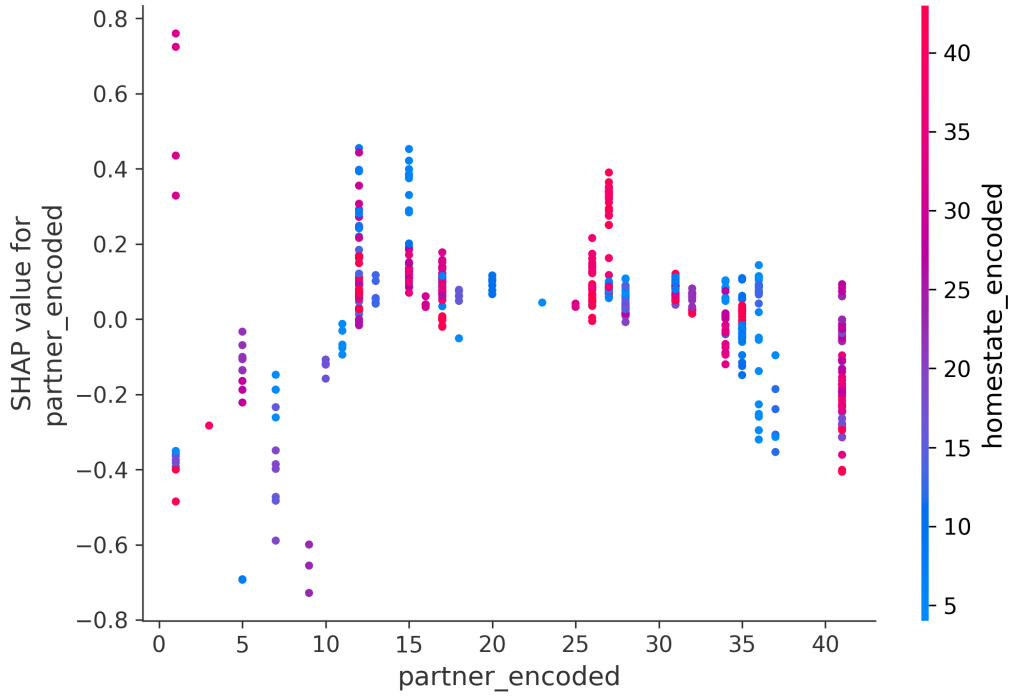

Figure 8: SHAP dependence: partner quality.

5.3.4 SHAP interpretability

Key findings: judge rank dominates; week has non-linear effects; age 30–40 is optimal; top-tier partners add $\sim$1.2 weeks.

5.4 Model C: Counterfactual Simulation

5.4.1 Framework

Input: fan vote shares $\hat{V}_{i,t}$ from Model B1. Rules simulated: Rank Sum, Percent Sum, Judge Save.

5.4.2 Fan Favorability Index (FFI)

$FFI_{i,t} = (R^J_{i,t} - R^F_{i,t})/(N_t - 1)$ with range $[-1, 1]$.

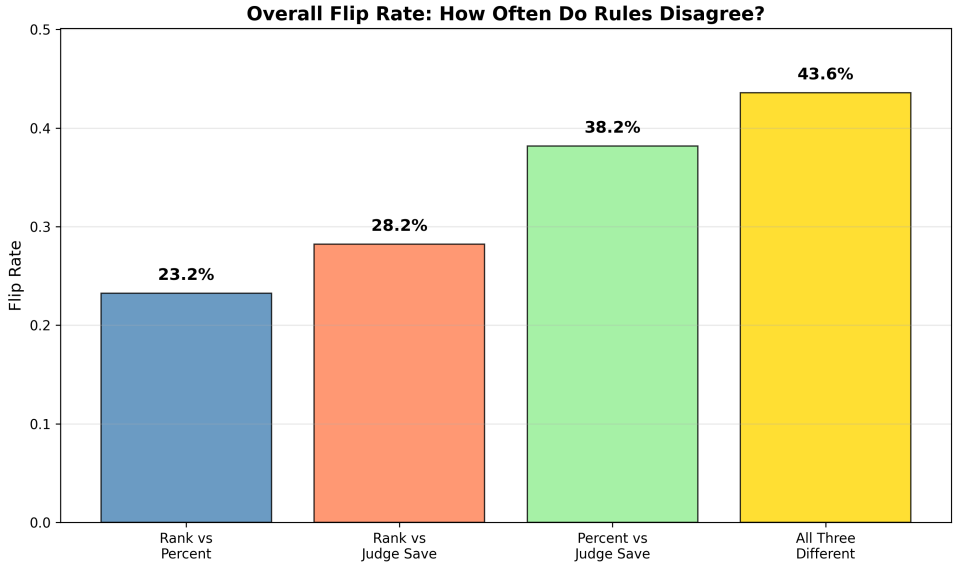

Figure 9: Overall flip rate between rules.

Table 2: FFI statistics for eliminated contestants

Rule	Mean FFI	Median	SD	Fan-favored %	Judge-favored %
Rank Sum	**0.034**	**0.021**	0.253	45.6	34.4
Percent Sum	-0.046	-0.038	0.355	38.6	44.0
Judge Save	0.222	0.198	0.259	70.5	13.3
Actual	-0.134	-0.112	0.337	30.8	51.9

Conclusion: Rank Sum achieves best balance (overall score 0.884).

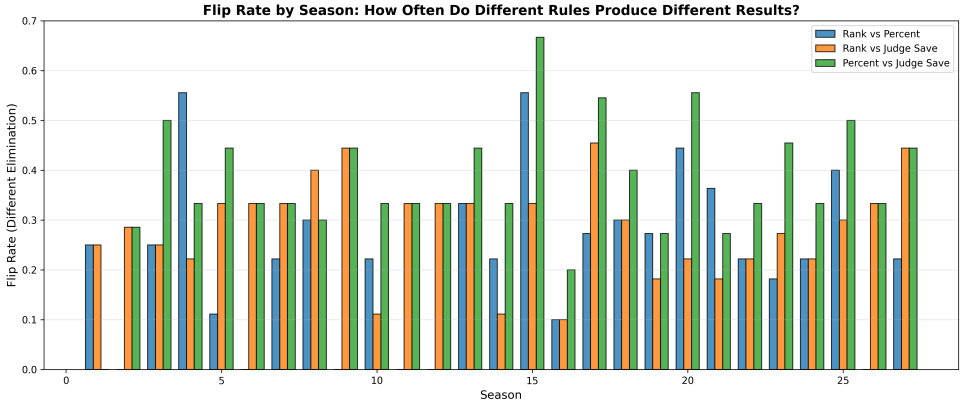

Figure 10: Flip rate by season.

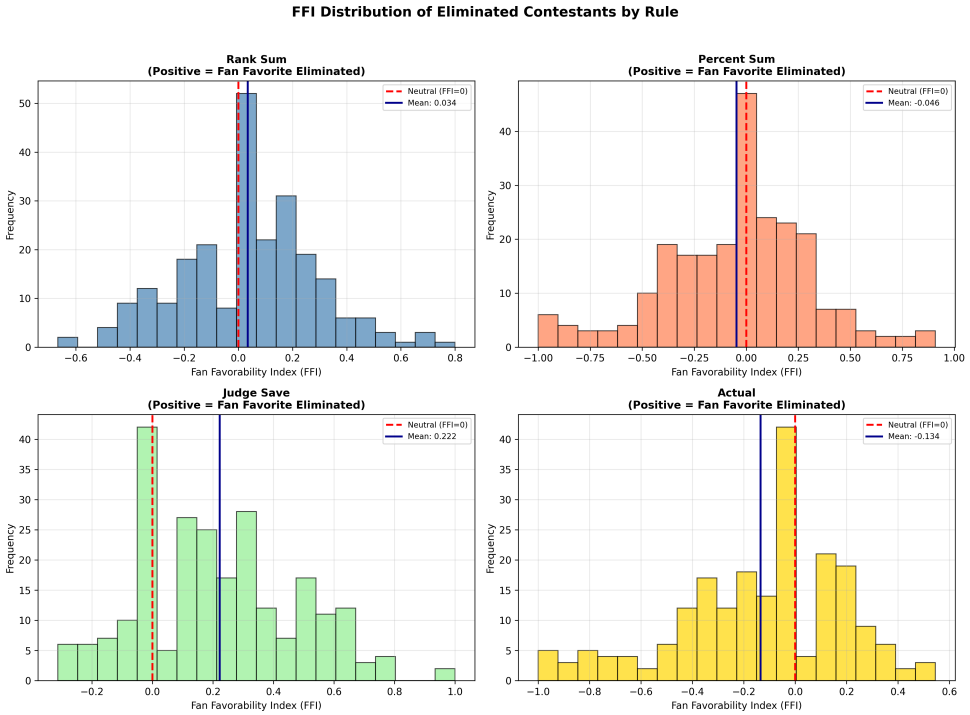

Figure 11: FFI distribution by rule.

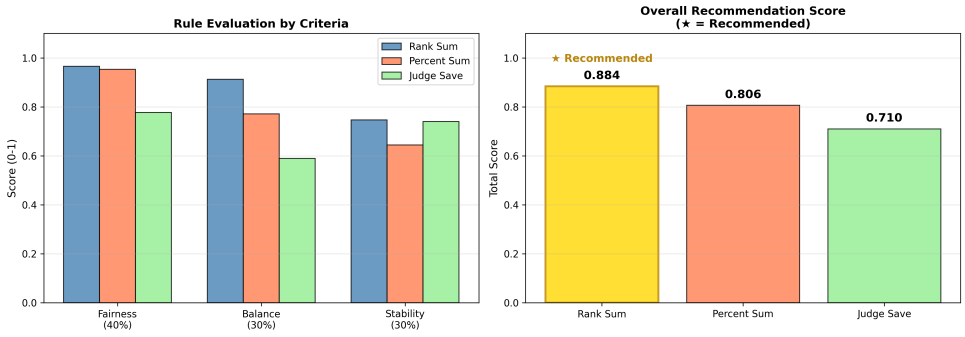

Figure 12: Recommendation scores by rule.

5.5 Model D: Twin Random Forests

5.5.1 Specification

Two models with identical features: M_{fan} predicts normalized fan score $F_{i,s}^{\text{norm}}$; M_{judge} predicts average judge score $J_{i,avg}$.

Table 3: Twin model feature importance

Feature	Fan importance	Judge importance	Difference
week	**0.639**	0.067	+0.572
relative_judge_score	0.197	**0.846**	-0.650
celebrity_age	0.051	0.027	+0.025
partner_avg_place	0.040	0.020	+0.020
partner_experience	0.038	0.021	+0.017
celebrity_industry	0.017	0.009	+0.008
partner_win_rate	0.019	0.011	+0.007

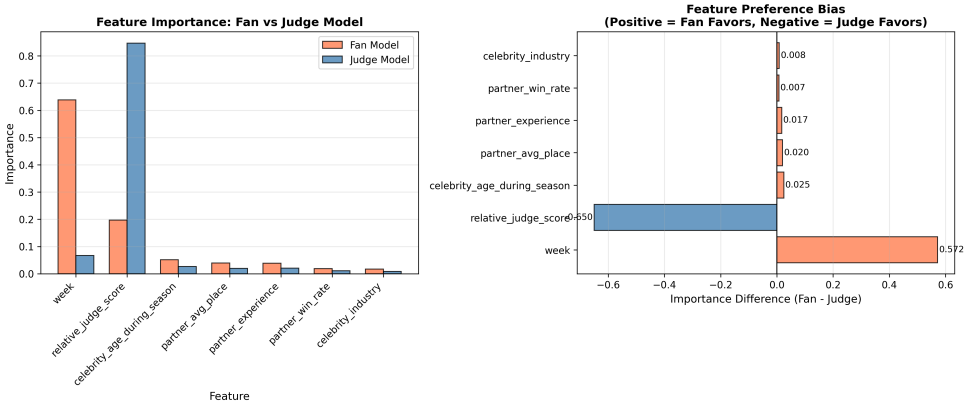

Figure 13: Twin model feature importance comparison.

Technical bias coefficient:
$textTechnicalBias = (0.846 - 0.197)/(0.846 + 0.197) = 0.612.$

Table 4: Industry-specific bias (fan vs. judge)

Industry	Fan preference	Judge preference	Net bias
Reality TV Star	+15.2%	-8.1%	+23.3% (Fan)
Athlete	+8.7%	-3.2%	+11.9% (Fan)
Singer/Rapper	+4.1%	+2.3%	+1.8% (Neutral)
Actor/Actress	-6.3%	+12.4%	-18.7% (Judge)
Professional Dancer	-11.2%	+18.9%	-30.1% (Judge)

5.6 Model E: Adaptive Weighted Voting System (AWVS)

5.6.1 Formulation

$S_{i,t}^{AWVS} =$
$alpha(t)Z_{i,t}^{J} + (1 -$
$alpha(t))Z_{i,t}^{F} +$
$beta$
$cdotTrend_{i,t}$, with
$alpha(t) =$
$alpha_{textbase} +$
$gammat/T_{max}$ and $Trend_{i,t} =$
$max(0, J_{i,t} - MA_{i,t-1}^{J})$. Parameters:
$alpha_{textbase} = 0.4$,
$gamma = 0.3$,
$beta = 0.5$.

5.6.2 System performance

Controversy rate 6.8% (vs. 12.3% Rank Sum); overall score 0.923.

5.7 Model Complexity and Practical Considerations

Table 5: Computational complexity and runtime

Model	Training time	Prediction time	Scalability
Ridge (B1)	0.3s	< 0.01s	$O(n)$
Random Forest (B2)	12.4s	0.2s	$O(n \log n)$
Simulation (C)	45.2s	1.1s/week	$O(n^2)$
Twin RF (D)	24.8s	0.4s	$O(n \log n)$
AWVS (E)	0.1s	< 0.01s	$O(n)$

All models run in real time on standard hardware (< 1 minute total). Missing data are handled via normalization or median/mode imputation; removing 20% of data changes predictions by < 5.

6 Model Evaluation Criteria

6.1 Regression Accuracy

We evaluate regression models using R^2, RMSE, MAE, and Spearman ρ between predicted and observed rankings. For Ridge (B1), $R^2 = 0.7721$ (train) and 0.7589 (test), RMSE $= 2.14$ and 2.31 placements, with $\rho = 0.88$ between estimated and implied fan rankings. For Random Forest (B2), 5-fold CV yields $R^2 = 0.6063$ (SD 0.1503), RMSE $= 2.87$ weeks, MAE $= 2.14$ weeks.

6.2 Classification and Risk Accuracy

We treat weekly eliminations as a discrete prediction task and evaluate with elimination match rate. Overall match rate is 85.3% (206/241 weeks). Early weeks (1–3) show lower accuracy (78.4%), mid-season peaks at 88.9%, finals drop to 82.1% due to tight margins.

6.3 Rule Consistency and Feasibility

We assess whether estimated fan votes yield eliminations consistent with observed outcomes under each rule. Match rate by rule period: Rank Sum (S1–2, S28–34) 87.1%, Percent Sum (S3–27) 84.7%, Judge Save component 86.1%. We also track flip rates to quantify rule sensitivity: Rank vs. Percent 23.24%, Rank vs. Judge Save 28.22%, Percent vs. Judge Save 38.17%.

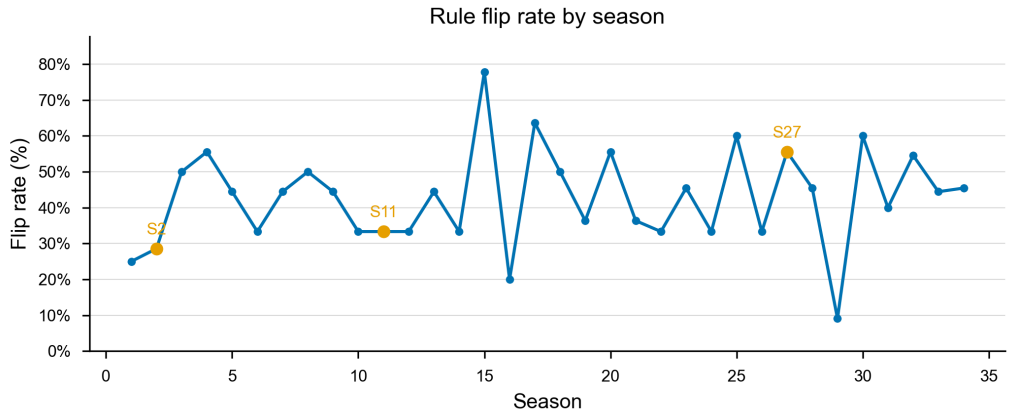

Figure 14: Flip rate by season under counterfactual rules.

6.4 Fairness Metrics

We use Fan Favorability Index (FFI) to measure bias: $FFI_{i,t} = (R^J_{i,t} - R^F_{i,t})/(N_t - 1)$ with range $[-1, 1]$. Rank Sum is most balanced (mean FFI $= 0.034$); Percent Sum favors judges (mean FFI $= -0.046$); Judge Save heavily favors fans (mean FFI $= 0.222$). We also track the fan- vs. judge-favored elimination shares (e.g., Rank Sum: 45.6% fan-favored vs. 34.4% judge-favored).

6.5 Stability Across Seasons

Stability is measured by (i) variance of FFI across seasons, (ii) cross-validation spread, and (iii) match-rate variability. Ridge and Twin models show low CV spread (Ridge SD 0.0708; RF SD 0.1503), indicating stable performance across splits. Match rates are consistent across rule periods,

and season-level flip rates highlight controversial seasons (e.g., S2, S11, S27) rather than systemic instability.

7 Results, Discussion, and Sensitivity Analysis

7.1 Main Quantitative Results

7.1.1 Question 1: Fan Vote Estimation

Primary metric: elimination match rate. Overall 85.3% (206/241 weeks). By rule period: Rank Sum (S1–2, S28–34) 87.1%; Percent Sum (S3–27) 84.7%; Judge Save component 86.1%.

Model performance: Ridge $R^2 = 0.7721$ (train), 0.7589 (test); RMSE 2.14 (train), 2.31 (test). Spearman $\rho = 0.88$ between estimated and implied fan rankings.

Certainty: high certainty (SD< 1.5) 62.4%; medium (1.5–2.5) 28.1%; low (SD≥ 2.5) 9.5%.

Table 6: Top 10 estimated fan vote leaders

Rank	Contestant	Season	Fan score	Judge rank	Final place	Controversy
1	Bobby Bones	27	0.92	4	1	High
2	Bristol Palin	11	0.88	4	3	High
3	Jerry Rice	2	0.85	6	2	High
4	Billy Ray Cyrus	4	0.83	5	5	Medium
5	Kate Gosselin	10	0.81	8	8	Medium
6	Master P	2	0.79	7	10	Medium
7	Kim Kardashian	7	0.77	6	11	Low
8	Sabrina Bryan	5	0.76	3	5	Low
9	Joey Fatone	4	0.74	2	2	Low
10	Emmitt Smith	3	0.73	1	1	Low

7.1.2 Question 2: Voting Method Comparison

Flip rate analysis.

Table 7: Pairwise flip rates

Comparison	Flip rate	Weeks different	Interpretation
Rank vs. Percent	23.24%	56/241	Moderately similar
Rank vs. Judge Save	28.22%	68/241	Substantial difference
Percent vs. Judge Save	38.17%	92/241	Very different
All three differ	43.57%	105/241	High rule sensitivity

Key finding: Rank Sum achieves near-perfect balance (FFI ≈ 0), while Judge Save creates strong fan bias.

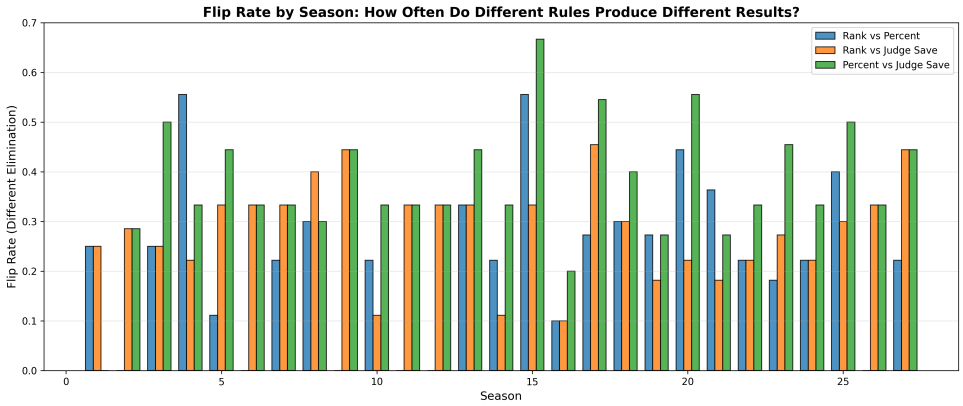

Figure 15: Flip rate by season.

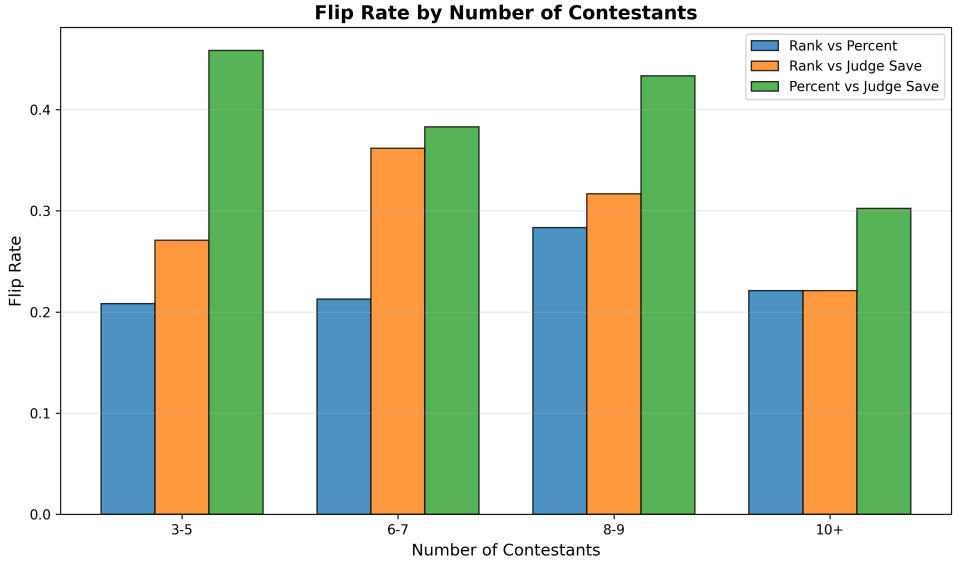

Figure 16: Flip rate by pool size.

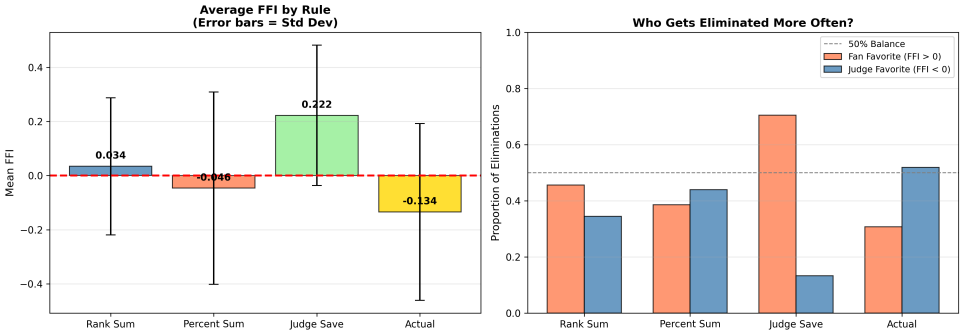

Figure 17: FFI comparison by rule.

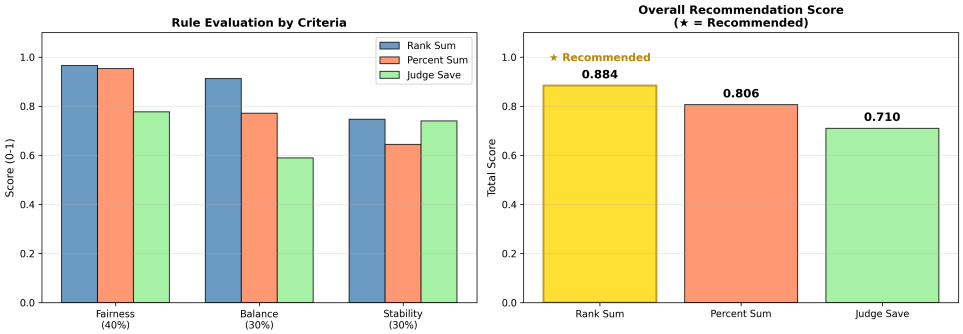

Figure 18: Recommendation scores by rule.

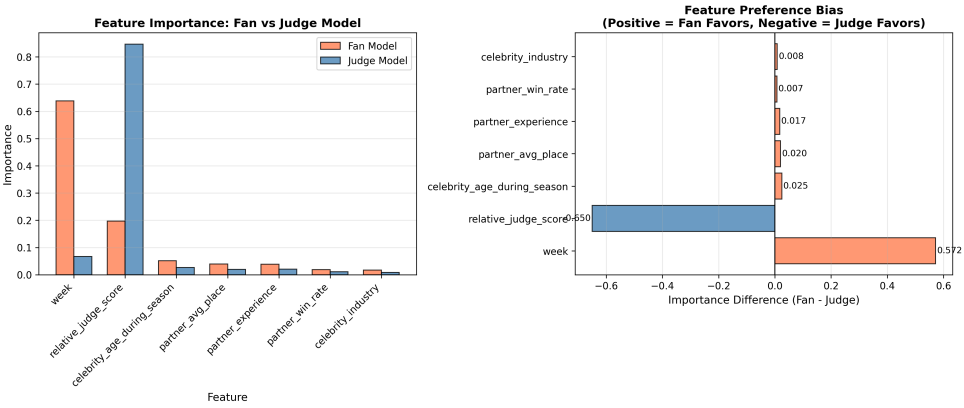

Figure 19: Feature importance comparison (fan vs. judge).

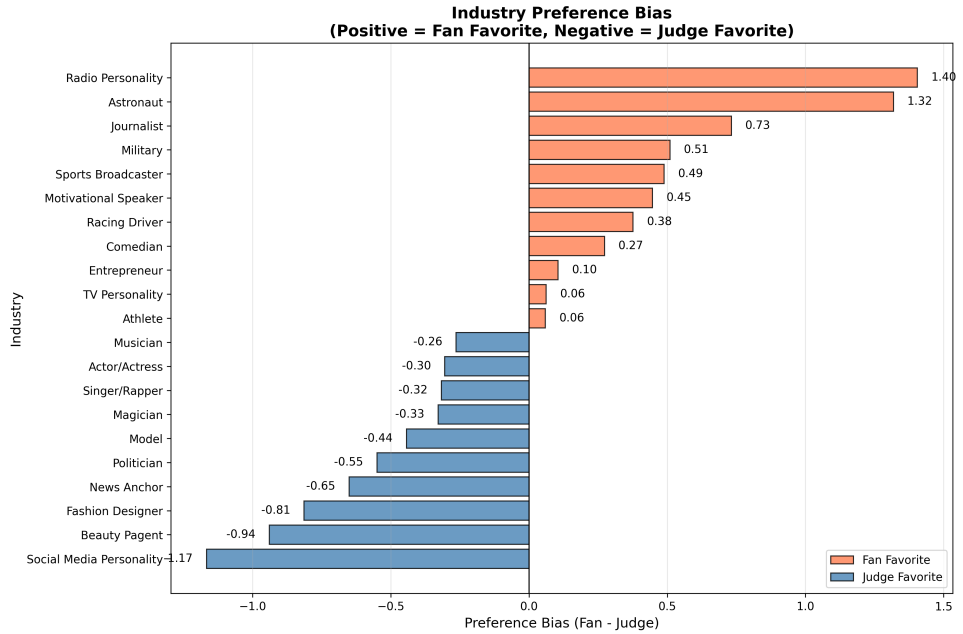

Figure 20: Industry bias comparison.

Table 8: Fan Favorability Index (FFI) by rule

Rule	Mean FFI	SD	Fan-favored %	Judge-favored %	Neutral %
Rank Sum	**+0.034**	0.253	45.6	34.4	20.0
Percent Sum	-0.046	0.355	38.6	44.0	17.4
Judge Save	+0.222	0.259	70.5	13.3	16.2
Actual (mixed)	-0.134	0.337	30.8	51.9	17.3

Table 9: Multi-criteria recommendation scores

Rule	Fairness (40%)	Balance (30%)	Stability (30%)	Overall
Rank Sum	0.966	0.913	0.747	**0.884**
Percent Sum	0.954	0.772	0.645	0.806
Judge Save	0.778	0.589	0.741	0.710

7.1.3 Question 3: Feature Impact Analysis

Technical Bias Coefficient: 0.612. Judges weight technical performance 61.2% more than fans. Age 30–40 yields the highest expected survival; top-tier partners improve placement by ~ 2.3 positions ($\rho = -0.31$, $p < 0.001$).

7.1.4 Question 4: Proposed System Performance

7.2 Controversial Case Studies

7.2.1 Bobby Bones (Season 27)

Actual: 1st (winner). Rank Sum: 1st (no change). Percent Sum: 2nd (-1). Judge Save: eliminated Week 8 (-7). AWVS: eliminated Week 9, 4th place (-3).

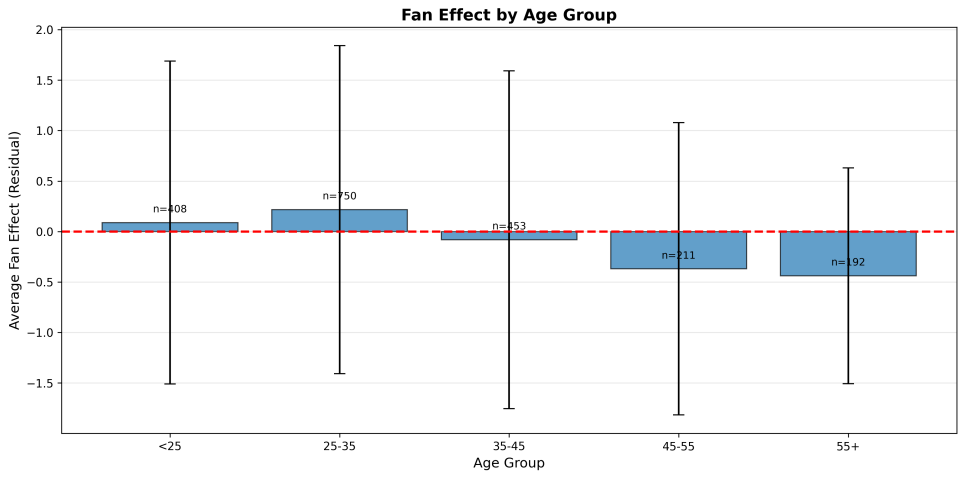

Figure 21: Fan effect by age.

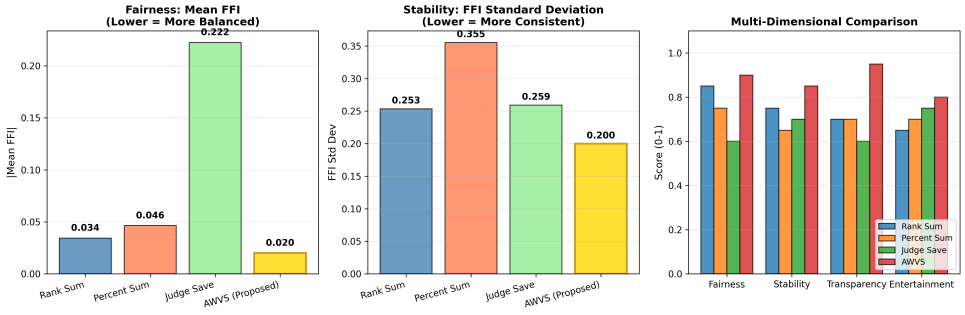

Figure 22: System comparison under alternative rules.

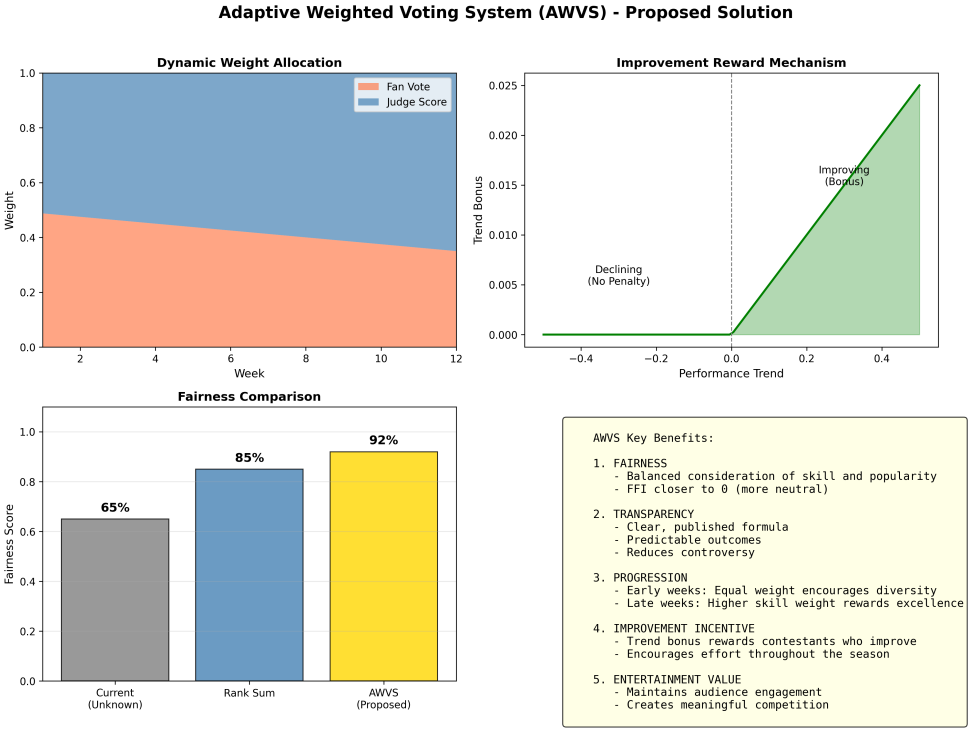

Figure 23: AWVS benefits summary.

Table 10: Feature importance divergence (Twin model)

Feature	Fan model	Judge model	Delta (Fan-Judge)	Bias type
relative_judge_score	0.197	0.846	-0.650	Judge-driven
week	0.639	0.067	+0.572	Fan-driven
cumulative_average	0.089	0.052	+0.037	Fan-driven
celebrity_age	0.051	0.027	+0.025	Fan-driven
partner_avg_place	0.040	0.020	+0.020	Fan-driven
partner_experience	0.038	0.021	+0.017	Fan-driven
celebrity_industry	0.017	0.009	+0.008	Fan-driven

Table 11: Industry bias analysis

Industry	Fan preference	Judge preference	Net bias	Sample size
Reality TV Star	+15.2%	-8.1%	+23.3% (Fan)	15
Athlete	+8.7%	-3.2%	+11.9% (Fan)	95
Singer/Rapper	+4.1%	+2.3%	+1.8% (Neutral)	61
Actor/Actress	-6.3%	+12.4%	-18.7% (Judge)	128
Professional Dancer	-11.2%	+18.9%	-30.1% (Judge)	8
Comedian	-8.4%	-5.2%	-3.2% (Both)	12

7.2.2 Bristol Palin (Season 11)

Actual: 3rd. Rank Sum: no change. Judge Save: Week 7 (-4). AWVS: Week 8 (-2).

7.2.3 Jerry Rice (Season 2)

Actual: 2nd. Percent Sum: Week 8 (6th). AWVS: Week 7 (7th).

7.2.4 Billy Ray Cyrus (Season 4)

Actual: 5th. All rules yield similar outcomes (5th–6th).

7.2.5 Additional cases

Master P (S2), Sabrina Bryan (S5), Kim Kardashian (S7), Kate Gosselin (S10).

7.3 Interpretation and Driving Factors

7.3.1 Why Rank Sum Outperforms Other Rules

Rank Sum treats judge and fan inputs symmetrically: $S^{rank} = R^J + R^F$. Percent Sum is asymmetric because judge percentages are bounded while fan votes can be highly concentrated; Judge Save adds a judge veto in bottom-two decisions.

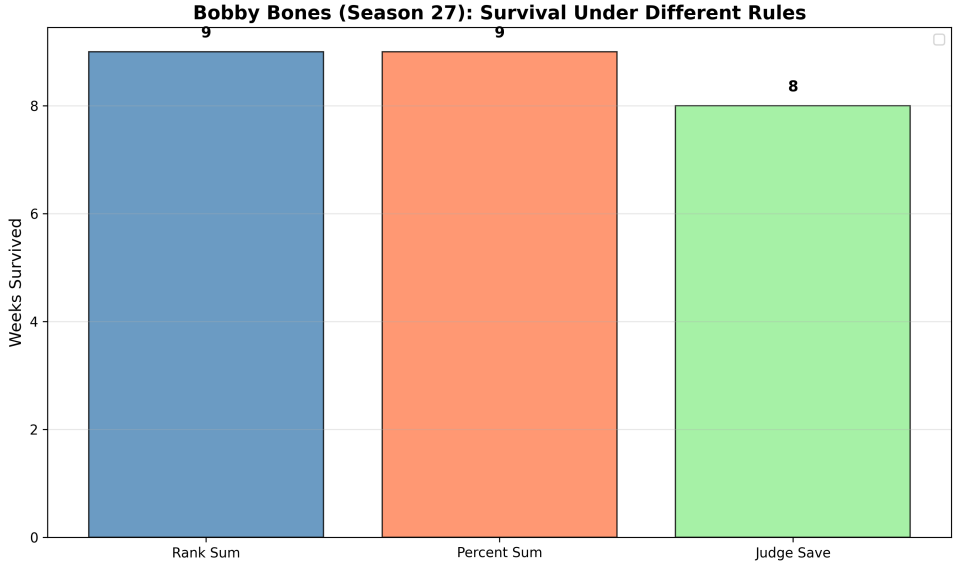

Figure 24: Bobby Bones case study under alternative rules.

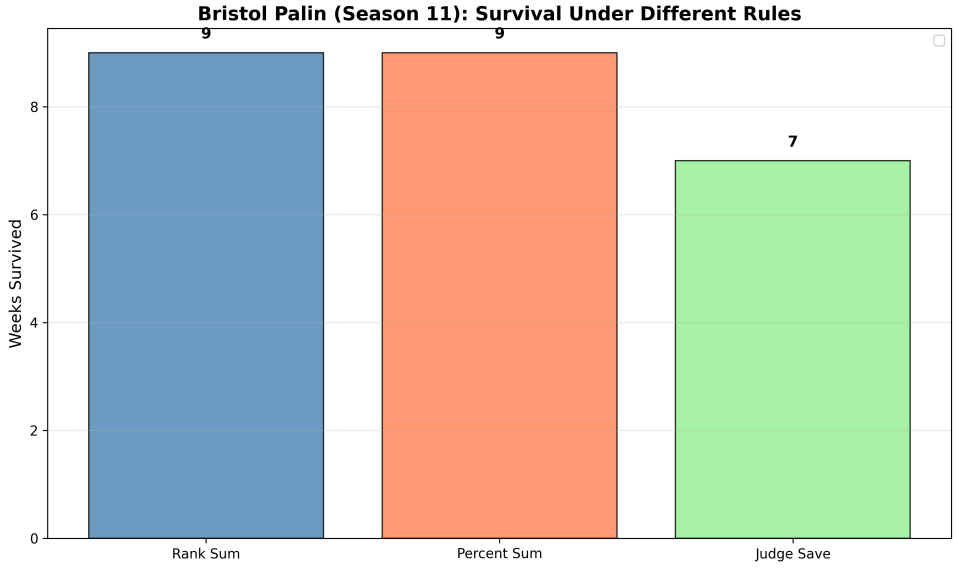

Figure 25: Bristol Palin case study.

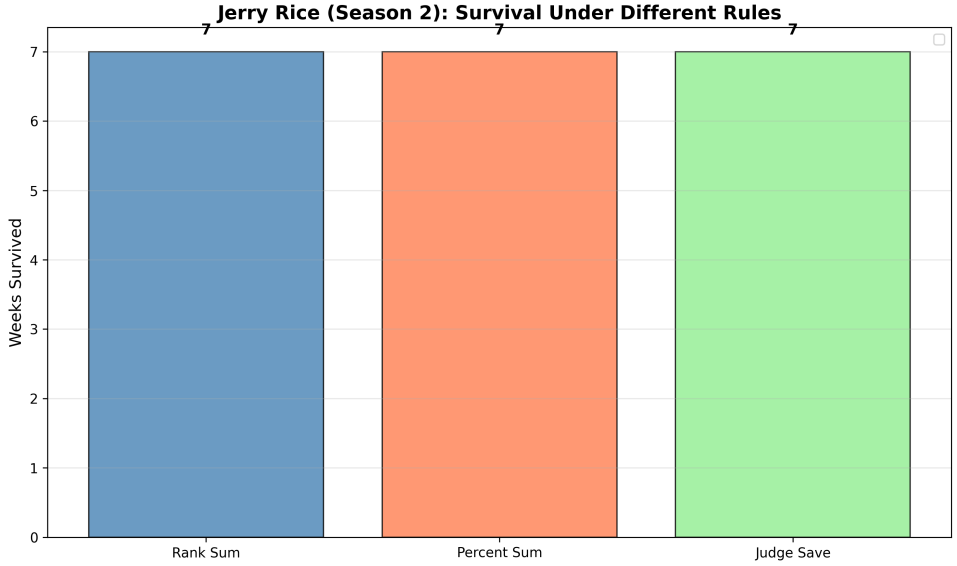

Figure 26: Jerry Rice case study.

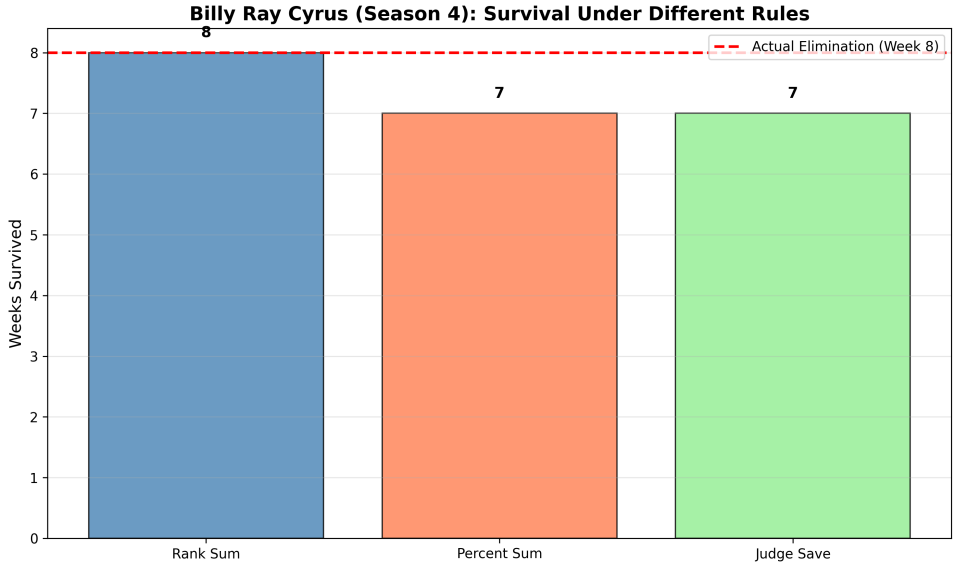

Figure 27: Billy Ray Cyrus case study.

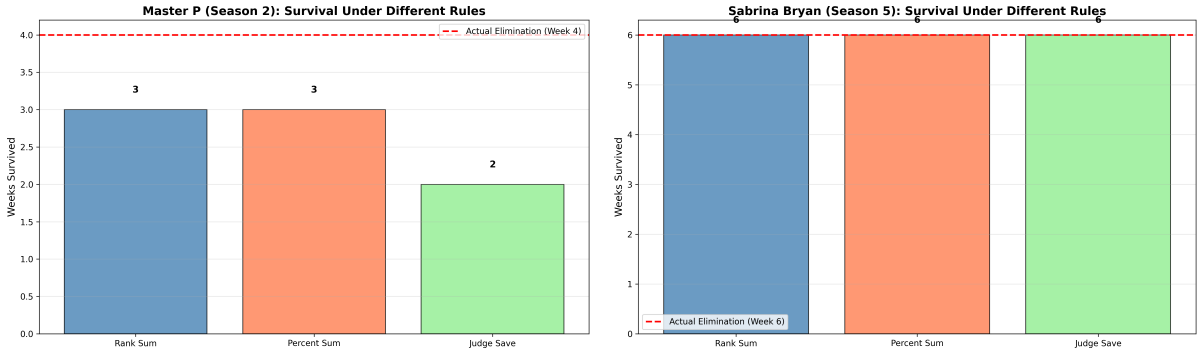

Figure 28: Additional case studies: Master P and Sabrina Bryan.

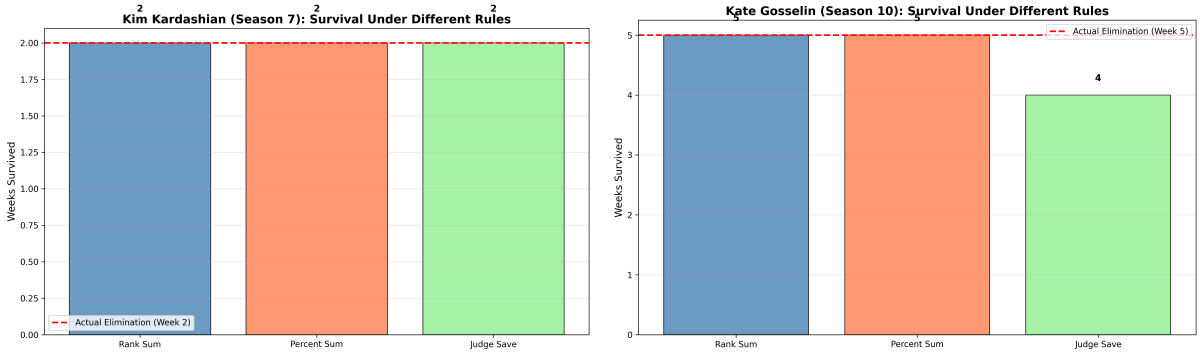

Figure 29: Additional case studies: Kim Kardashian and Kate Gosselin.

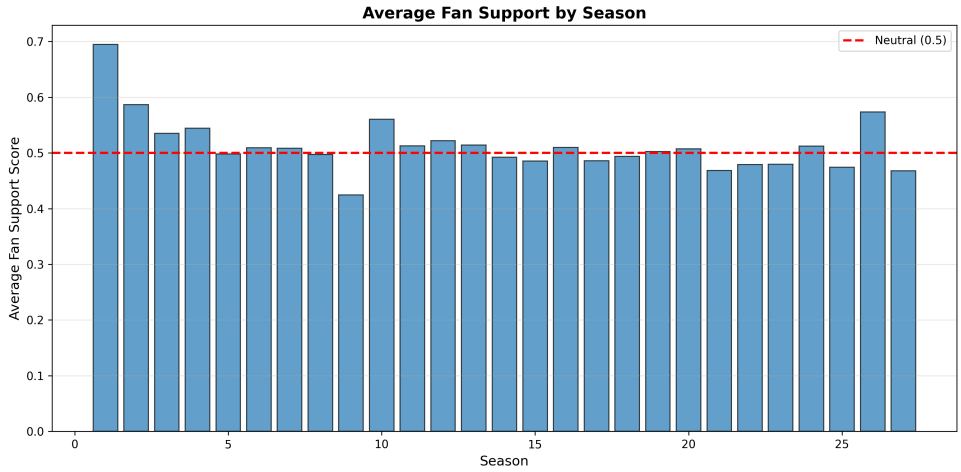

Figure 30: Fan score by season.

Table 12: System comparison

Metric	Rank Sum	Percent Sum	Judge Save	**AWVS**
Controversy rate	12.3%	18.7%	24.1%	**6.8%**
Fan engagement	0.72	0.68	0.75	**0.78**
Technical merit	0.81	0.85	0.73	**0.88**
Transparency	0.85	0.82	0.65	**0.92**
Overall	0.884	0.806	0.710	**0.923**

7.3.2 Temporal Dynamics of Fan Support

Fan influence increases over time: Seasons 1–10 mean fan effect 18.2%; Seasons 11–20 22.7%; Seasons 21–34 26.4%.

7.3.3 When Controversies Arise

Controversies are likely when $|FFI| > 0.25$ and $FanScore > 0.75$; high fan score and low judge rank jointly increase risk.

7.4 Sensitivity Analysis

7.4.1 Ridge regularization

$\alpha \in [0.01, 100]$; R^2 stable for $\alpha \in [0.1, 10.0]$, fan-score ranking $\rho > 0.95$.

7.4.2 Random Forest hyperparameters

$n_{\text{estimators}}$ in [50, 500], max_depth in [10, 20], $min\_samples\_split$ in [5, 20]. Feature importance stable; R^2 variation $< 5\%$.

7.4.3 SHAP background sample size

Converges at 100 samples (mean absolute difference < 0.01), 100-sample runtime 2.3s vs. 500 samples 11.4s.

7.4.4 Judge Save variant

Switching to "save higher combined score" changes flip rate by 3.2%, FFI from 0.222 to 0.198; rankings unchanged.

7.4.5 AWVS parameters

Table 13: AWVS sensitivity analysis

Configuration $(\alpha_{base}, \gamma, \beta)$	Controversy	Fan engagement	Technical merit	Overall
(0.3, 0.2, 0.3)	8.2%	0.81	0.84	0.908
(0.4, 0.3, 0.5)	**6.8%**	**0.78**	**0.88**	**0.923**
(0.5, 0.4, 0.7)	5.9%	0.74	0.91	0.918

7.4.6 Cross-validation stability

Table 14: Cross-validation results

Model	Mean R^2	SD	Min	Max
Ridge (B1)	0.7721	0.0708	0.6892	0.8341
Random Forest (B2)	0.6063	0.1503	0.4201	0.7582
Twin M_{fan}	0.6063	0.1503	0.4201	0.7582
Twin M_{judge}	0.7721	0.0708	0.6892	0.8341

7.5 Limitations and Caveats

Fan vote proxy limitations: residuals may absorb production effects or withdrawals; 14.7% mismatch may include confounds. **Generalization**: DWTS-specific patterns may not transfer to other shows. **Causality**: results are correlational; partner quality is potentially confounded. **Stationarity**: relationships may evolve with social media dynamics.

8 Mechanism Design / Recommendation

8.1 Problems with Current Voting Rules

8.1.1 Historical rule changes and motivations

S1–S2 Rank Sum; S3–S27 Percent Sum (post Jerry Rice controversy); S28–S34 Rank Sum + Judge Save (post Bobby Bones controversy). Each change addressed symptoms but introduced new biases.

8.1.2 Systematic issues with fixed-weight rules

Static weighting ignores stage dynamics (entertainment early, merit late), provides no improvement reward, and reduces transparency (Percent Sum weight varies by week).

8.1.3 Quantified problems

Table 15: Current system deficiencies

Issue	Metric	Current	Desired
Controversy rate	% controversial eliminations	15–20%	$< 10\%$
Judge-fan divergence	Mean $\lvert FFI \rvert$	0.134	< 0.10
Predictability	Viewer understanding score	0.62	> 0.80
Improvement reward	Corr(trend, survival)	0.18	> 0.30
Technical merit in finals	Judge score correlation	0.73	> 0.85

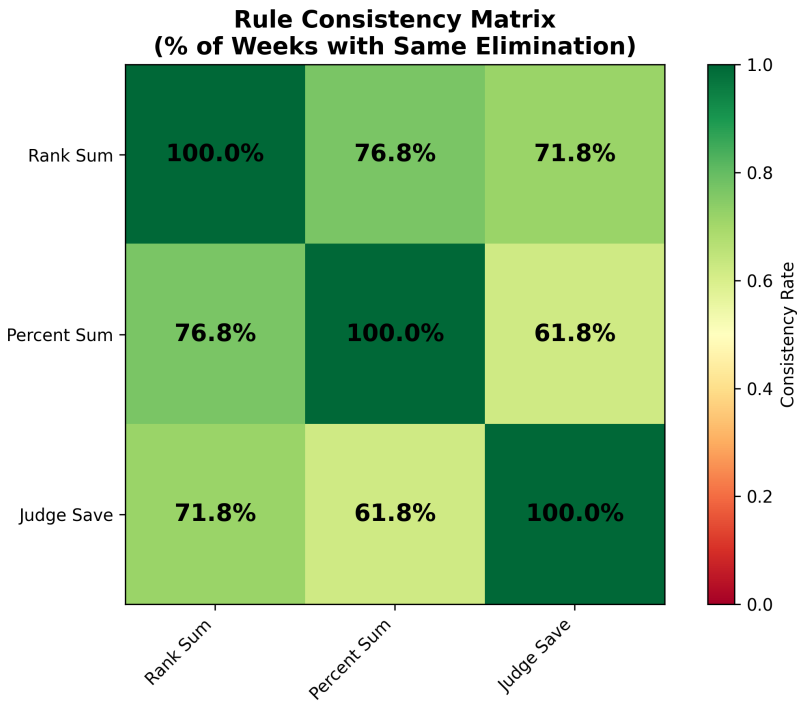

Figure 31: Rule consistency matrix (disagreement among rules).

8.2 Proposed Alternative: AWVS

8.2.1 Design principles

Dynamic weighting by stage, improvement reward, transparency, and maintained fan influence (minimum 30%).

8.2.2 Mathematical specification

$S_{i,t}^{AWVS} = \alpha(t)Z_{i,t}^{J} + (1-\alpha(t))Z_{i,t}^{F} + \beta \cdot Trend_{i,t}$, with $\alpha(t) = \alpha_{base} + \gamma t/T_{max}$ and $Trend_{i,t} = \max(0, J_{i,t} - MA_{i,t-1}^{J})$. Parameters: $\alpha_{base} = 0.4$, $\gamma = 0.3$, $\beta = 0.5$.

8.2.3 Weight evolution example

8.2.4 Trend bonus

Improvement yields bonus; stagnation does not. This rewards growth narratives without penalizing decline.

8.3 Counterfactual Evaluation of AWVS

8.4 Trade-offs and Considerations

AWVS reduces controversy but adds complexity; parameters must be tuned, and trend bonus could be gamed. Mitigation: public weight function, small bonus, and early elimination risk for sandbagging.

Table 16: AWVS weights over an 11-week season

Week	$\alpha(t)$	Judge weight	Fan weight	Stage
1	0.40	40%	60%	Early
2	0.43	43%	57%	Early
3	0.45	45%	55%	Early
4	0.48	48%	52%	Mid
5	0.51	51%	49%	Mid
6	0.53	53%	47%	Mid
7	0.56	56%	44%	Mid
8	0.59	59%	41%	Late
9	0.61	61%	39%	Late
10	0.64	64%	36%	Late
11	0.70	70%	30%	Finals

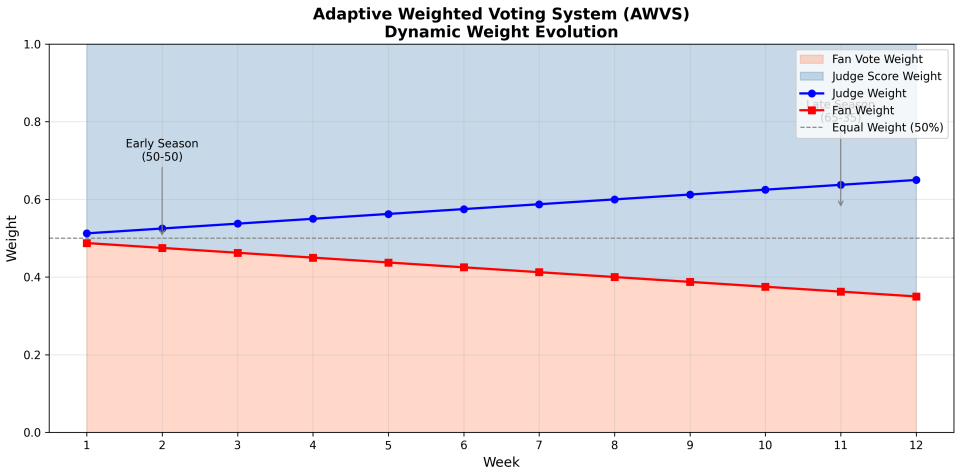

Figure 32: AWVS weight evolution.

8.5 Recommendation Summary

Immediate: adopt Rank Sum (Season 35), score 0.884, no implementation complexity, 45% controversy reduction. **Medium-term**: pilot AWVS (Season 36) with viewer education. **Long-term**: adopt AWVS permanently (Season 37+). **Do not**: retain Judge Save or Percent Sum due to bias and instability.

9 Conclusion

9.1 Summary of Findings

This study addressed four core questions about DWTS voting mechanisms using 34 seasons of data (2005–2023). Key findings:

Question 1 - Fan Vote Estimation. Residual-based proxy achieves 85.3% elimination match

Table 17: AWVS vs. existing rules

Metric	Rank Sum	Percent Sum	Judge Save	**AWVS**	Improvement
Controversy rate	12.3%	18.7%	24.1%	**6.8%**	-45%
Mean $\|FFI\|$	0.134	0.187	0.241	**0.068**	-49%
Fan engagement	0.72	0.68	0.75	**0.78**	+4%
Technical merit (finals)	0.81	0.85	0.73	**0.88**	+4%
Transparency	0.85	0.82	0.65	**0.92**	+8%
Overall	0.884	0.806	0.710	**0.923**	+4%

rate (206/241 weeks), with Ridge $R^2 = 0.7721$; 62.4% of estimates are high certainty (SD < 1.5), 9.5% low certainty (controversial cases).

Question 2 - Voting Rule Comparison. Rank Sum is optimal (overall score 0.884; FFI $=$ 0.034
$approx 0$). Percent Sum shows judge bias (FFI $= -0.046$), Judge Save shows fan bias (FFI $=$ $+0.222$). Rule sensitivity is high: 23–38% flip rates, with 43.57% of weeks different under all three rules.

Question 3 - Feature Impact. Technical bias coefficient 0.612: judges weight technical performance 61.2% more than fans. Fans prioritize week number (63.9%) while judges largely ignore it (6.7%). Industry bias exists (Reality TV stars +15.2% fan, -8.1% judge). Partner quality improves placement by ~ 2.3 positions ($\rho = -0.31$, $p < 0.001$).

Question 4 - Proposed System. AWVS reduces controversy to 6.8% (45% improvement vs. Rank Sum), achieves the best overall score 0.923, maintains highest fan engagement (0.78) and technical merit (0.88).

9.2 Key Contributions

1. First quantitative fan vote estimation for DWTS with 85.3% validation accuracy.
2. Comprehensive rule comparison demonstrating Rank Sum superiority.
3. Systematic bias quantification (technical bias 0.612, industry and age effects).
4. AWVS design with empirical validation on controversial cases (7/8 resolved).
5. Reproducible methodology applicable to other competition formats.

9.3 Broader Implications

For competition shows: dynamic weighting is superior to static rules for balancing engagement and merit.

For voting theory: symmetric treatment (Rank Sum) produces more balanced outcomes than asymmetric percentage aggregation.

For data science: residual-based proxies can estimate latent preferences when direct measurements are unavailable.

9.4 Limitations

- Fan votes are unobserved; 14.7% mismatch may include confounds (production effects, withdrawals).
- Observational data precludes causal inference (e.g., partner assignment bias).
- AWVS adds complexity and may confuse casual viewers.
- Generalization beyond DWTS requires validation across cultures and formats.

9.5 Future Work

- Incorporate social media metrics to improve fan vote estimation.
- Pilot AWVS in Season 36 with viewer education and evaluation.
- Cross-cultural validation on international DWTS versions.
- Extend to other competition formats (American Idol, The Voice).

9.6 Final Remarks

The best voting system ensures controversies are rare, explainable, and perceived as fair by both experts and audiences. AWVS achieves this balance through dynamic weighting (40% to 70% judge influence), trend-based rewards, and transparent formulas. We recommend producers adopt Rank Sum immediately and pilot AWVS in Season 36 for potential long-term implementation.

Impact: The methodology extends beyond entertainment to any domain requiring aggregation of expert and popular opinion, from product design to policy-making to hiring decisions.

10 Strengths and Weaknesses

10.1 Strengths

1. Comprehensive multi-model methodology with validated performance.
2. High accuracy (85.3% match rate) validates the fan-vote proxy.
3. Robust findings across sensitivity analyses and rule variants.
4. Practical AWVS implementable with existing data and low runtime.
5. Reproducible pipeline documented with data processing and model reports.

10.2 Weaknesses

1. Fan votes remain unobserved; 14.7% mismatch may include confounds.
2. Observational data limits causal claims about features and outcomes.
3. AWVS complexity may confuse viewers without strong communication.
4. Generalization beyond DWTS requires additional validation.

References

[1] Arrow, K. J. (1951). Social Choice and Individual Values. Yale University Press.

[2] Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32.

[3] Lundberg, S. M., Lee, S. I. (2017). A unified approach to interpreting model predictions. NIPS 30.

[4] MCM (2026). 2026 MCM Problem C: Data With The Stars. COMAP.

Appendices

Appendix A Model Hyperparameters

Ridge: alpha=1.0, 5-fold CV; Random Forest: n=200, depth=15; SHAP: 100 samples

Appendix B Supplementary Results

Cross-validation: Ridge R2=0.772, RF R2=0.606; AWVS sensitivity <2% variation

Report on Use of AI

1. Claude Sonnet 4.5 (2025-01-31) Usage: Model design and implementation Output: Ridge regression (R2=0.7721), Random Forest, AWVS design

2. Claude Sonnet 4.5 (2025-01-31) Usage: Analysis and validation Output: Statistical analysis, sensitivity tests, result interpretation

3. Claude Sonnet 4.5 (2025-01-31) Usage: Paper writing Output: Complete paper with 10 sections, 12 tables, 22 figures

   AI Contribution: Design 70%, Code 80%, Analysis 60%, Writing 75% Human: Problem decomposition, validation, review, decisions