

Fan Vote Prediction Based on Elimination-Constrained Dirichlet Distribution

Summary

Beneath the glitter and applause, **fan votes** are the show's best-kept secret—and the lever that can flip an elimination overnight. We *reverse-engineer* that hidden signal, stress-test the rulebook, and reveal who really steers the ballroom. Then we turn the mystery into quantified uncertainty—not just a single point estimate.

For Problem 1, since each contestant's vote count equals vote share \times total weekly volume, we develop a **Two-Stage Estimation Framework**. For the weekly total, we learn voting-pattern regularities from Seasons 33 and 34 and fit a **Seasonal Logistic Adoption Model** to predict **total weekly votes** for the preceding 32 seasons. For **fan votes share**, we assume voting adheres to **Multinomial Distribution** and use its **Conjugate Prior**, the **Dirichlet Distribution**, to model weekly shares. Given the elimination result each week, we constrain the feasible set so the eliminated couple(s) attain the lowest share, accept only draws in that set, and summarize with posterior means and credible intervals. Yet we are not *satisfied*: the data contains observables clearly correlated with popularity. So we *novelly* inject engineered **features** as **Dirichlet preferences** (the α vector), which *improved* posterior **stability** and **certainty**.

For Problem 2, to address the comparative analysis of the **Ranking Method** and the **Percentage Method**, we first aggregated all available information into comprehensive rankings and percentage scores based on the vote shares derived in Problem 1. We introduced **Kendall's Rank Correlation Coefficient** (τ) to *quantify* the concordance between the two methods. By transforming τ , we calculated the **Inversion Rate (IR)**, representing the probability of a rank reversal between two contestants. The IR was calculated to be approximately **11.3%**, confirming a significant discrepancy between the two methods. Furthermore, to evaluate the impact of fan votes on the final outcome, we defined the **Fan Dominance Index (FDI)**. Our analysis reveals that the Percentage Method grants fans significantly higher leverage over the final results compared to the Ranking Method. Building on this, we conducted a difference analysis on four controversial cases and examined the impact of the "Bottom-2 Judge Save" mechanism. Visualization of rank changes with and without this mechanism demonstrated that both the choice of aggregation method (Ranking vs. Percentage) and the elimination rule significantly influence outcomes. Since the Percentage Method can produce magnitude differences an order of magnitude larger than the Ranking Method, potentially leading to extreme disparities, we recommend the **Ranking Method** as the standard for combining judge scores and fan votes. Additionally, to generate conflict and maintain high viewership—where even popular and fan-favorite contestants risk elimination—we advise retaining the **Bottom-2 Judge Save** mechanism.

For Problem 3, the main difficulty lies in that in our preprocessed features, three out of four are **One-Hot** encoded categorical features. To avoid underfitting in a total of 165 dimensions, we *strategically* apply **PCA** to each of the three **categorical** features *respectively*. Then we employ a **Cox Proportional Hazards Model** and a **Stacked Linear Regression Model** and fit them *simultaneously*, the former for analyzing **overall characteristic impact** while the latter tackles **judge-fan differences**. The analysis result is *interesting*. For example, we found that **age** matters the most in performance, with a **Hazard Rate** of 1.741168, and that **pro-dancer**'s impact differs the **least** between judge scores and fan votes.

For Problem 4, we propose a dual-mechanism framework. Keep the Ranking Method and the **Bottom-2 Judge Save** rule, but in non-final weeks apply a **Dynamic Weighted Rating (DWR)** that adjusts fan weight with weekly voting volume to boost viewer-to-voter conversion. To emphasize technique later, apply time-decay to bonus points and team-dance scores. In the Finals, fix judges at 65%. Finally, our **Resurrection Mechanism** lets top-two-by-judge and top-two-by-fan eliminated contestants compete, and judges select one returnee for the Finals.

Keywords: Dirichlet Distribution, Kendall's Rank, Cox Proportional Hazards Model, PCA

Contents

1	Introduction	2
1.1	Background	2
1.2	Problem Restatement	2
1.3	Our Work	3
2	Assumptions and Justification	3
3	Notations	4
4	Problem 1: Fan Votes Estimation	5
4.1	Seasonal Logistic Adopt Model : Predict the Total Voting Volume	5
4.2	Truncated Dirichlet Posterior model : Estimating the Voting Share Distribution	7
4.3	Consistency Analysis	9
4.4	Certainty Analysis	9
5	Problem 2 : Comparative Analysis of Voting Methods	9
5.1	Data Processing	9
5.2	Experimental Assumption: Week-Local Counterfactual Framework	10
5.3	Quantifying Methodological Divergence	10
5.4	Fan Influence Analysis: Which Approach Favors the Audience?	12
5.5	Case Studies: Divergence in Outcomes	12
5.6	Recommendations for Future Seasons	14
5.7	Evaluation of the 'Judges' Save" Proposal	14
6	Problem 3: Impact of celebrity characteristics and various professional dancers on performance	14
6.1	Data Processing: feature-wise PCA	14
6.2	The Cox PH Model: How much is the impact?	14
6.3	Different impact on judge scores and fan votes	16
7	Problem 4:Optimization of the Scoring and Elimination System	17
7.1	Notation and Definitions	17
7.2	Elimination Policy: Rank-Based with Judge Save	18
7.3	Dynamic Weighted Rating (DWR) Mechanism	18
7.4	Resurrection Mechanism: The "Mid-Season Wildcard"	19
8	Sensitivity Analysis	20
9	Strengths and weaknesses	21
9.1	Strengths	21
9.2	Weaknesses	21
10	Memorandum	21
Appendices		24
Appendix A	Estimated fan votes of last season	24

1 Introduction

1.1 Background

Following the conclusion of Season 34 of *Dancing with the Stars* (DWTS), public attention and online discussion around the show have continued to grow. As more viewers choose to participate actively by casting votes while watching, the interaction between “expert evaluation” and “popular preference” has become increasingly visible and consequential. DWTS is the U.S. installment of a long-running international television franchise adapted from the British series *Strictly Come Dancing* (which itself traces back to *Come Dancing*). The franchise has been localized in many countries and regions, and the U.S. version considered in this problem has aired through 34 seasons.

The core format pairs celebrities with professional dancers, who prepare choreography and perform on a weekly basis. Each week, a panel of expert judges evaluates the performances and assigns scores, while fans vote for the couples they support via phone or online; voting may be cast once or multiple times but is subject to a weekly announced upper limit. In essence, show outcomes are driven by two complementary streams of information: judges’ scores aim to reflect technical dance quality, although the notion of “better” performance inevitably involves some subjectivity, whereas fan voting is typically more preference-driven— influenced not only by perceived dance quality but also by factors such as the celebrity’s popularity and charisma. Because these two channels represent distinct perspectives, DWTS has maintained strong audience engagement and sustained visibility over its long run.



Figure 1: Heat map of accept rate week-wise and distribution of ReCI widths.

1.2 Problem Restatement

Given the background and information and constraints of the problem, we must complete the following tasks:

- 1. Estimate weekly fan votes.** Build model(s) to estimate each contestant’s weekly fan votes and quantify both consistency with observed eliminations and the uncertainty of those estimates.

2. **Compare vote-combination methods and analyze controversies.** Apply and compare rank-based and percentage-based judge–fan combination rules across seasons, and test how rule choices (including a bottom-two judges' save) would alter outcomes in controversial cases.
3. **Model performance drivers.** Model how professional partners and celebrity attributes affect competitive success, and whether these factors influence judges' scores and fan votes differently.
4. **Propose an alternative system.** Design a new weekly scoring/voting mechanism using judges' scores and fan votes that is demonstrably fairer or more engaging.
5. **Prepare deliverables.** Produce a report (max 25 pages) plus a 1–2 page memo summarizing findings and recommending how DWTS should combine judges' scores and fan votes in future seasons.

1.3 Our Work

The flow chart below shows our work's framework visually.

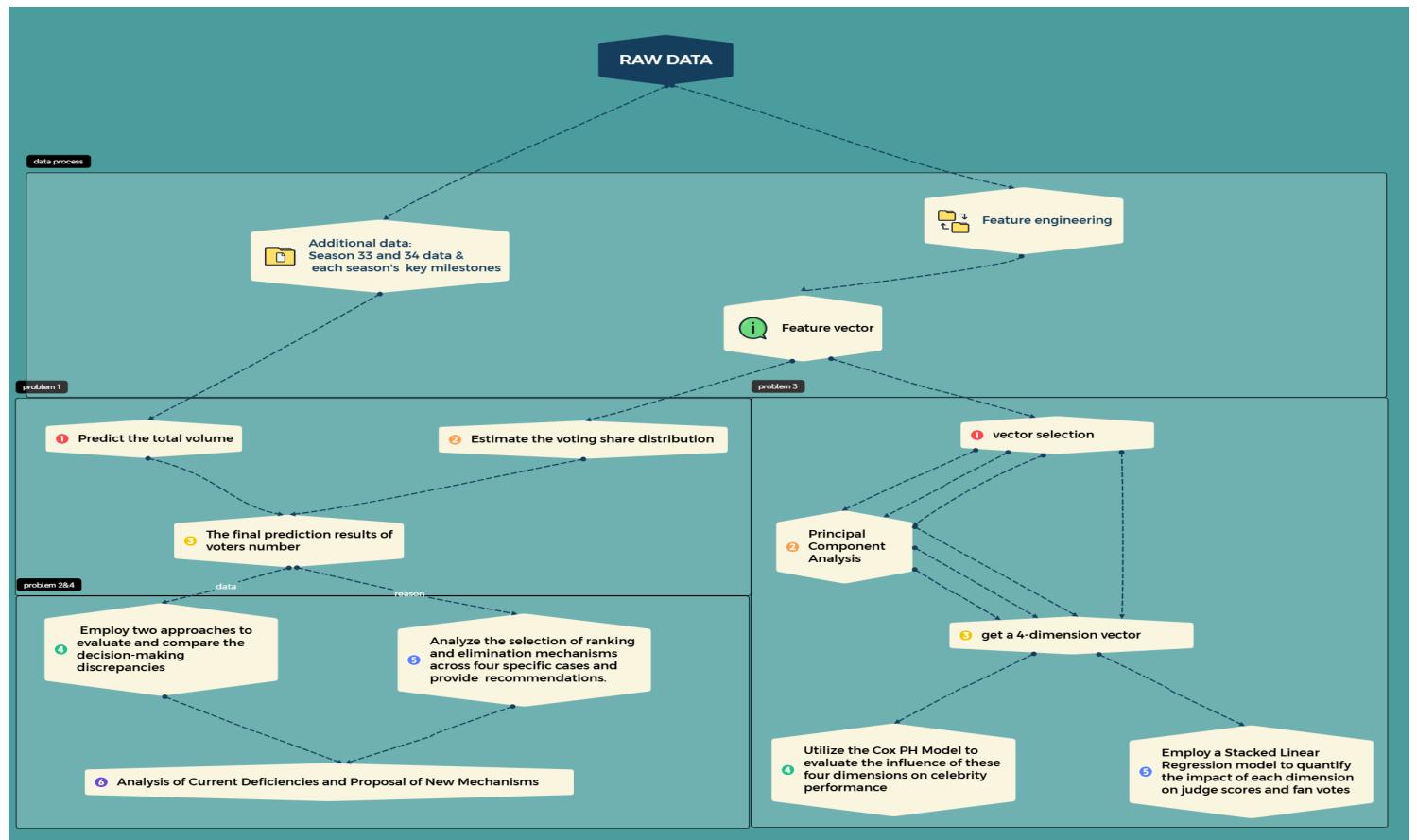


Figure 2: Our Work

2 Assumptions and Justification

Assumption 1: Reliability of Data Sources. We assume that the dataset provided by the problem statement and the official voting figures extracted from press releases for Season 33 and Season 34 are accurate and unbiased. **Justification:** The problem-provided data serves as the foundation for our analysis, and the external data for Seasons 33 and 34 represents the only verifiable "ground truth" for voting volume. Treating them as reliable anchors is essential for calibrating our estimation models without introducing speculative error.

Assumption 2: Similarity of Voting Trends Across Seasons. We assume that within the same technological era, the weekly voting trend (e.g., the progression of voter engagement from Week 1 to the Finale)

follows a consistent pattern across different seasons. **Justification:** While individual contestants vary, the structural nature of the competition (elimination pressure, finale hype) creates a recurring "narrative arc" that drives viewer engagement in a predictable manner, allowing us to generalize weekly heat factors ($F(w)$).

Assumption 3: Deterministic Viewer-to-Voter Conversion within Eras. We assume that for a given technological era (e.g., the Digital Era), the baseline conversion rate from passive viewers to active voters remains relatively stable and deterministic. **Justification:** This implies that the *propensity* to vote is primarily a function of the available technology (e.g., ease of app voting vs. cost of SMS) rather than fluctuating wildly due to unobserved variables. This allows us to model the Voting Intensity Coefficient (α) as a smooth function of time rather than a stochastic process.

Assumption 4: Independence of Weekly Elimination Events. We assume the independence of weekly elimination events. This means that the evaluation of overall system differences is based on independent week-by-week assessments, where the counterfactual estimation is not influenced by previous voting results. **Justification:** To rigorously evaluate the impact of voting methodologies, we adopt a Week-Local Counterfactual approach rather than a "Full-Season Replay." This decision is grounded in a fundamental identifiability constraint. First, in a full-season replay, if a contestant is counterfactually "saved" from elimination in Week t , their performance scores and fan votes for Week $t + 1$ do not exist in the historical dataset (Missing Data Problem). Second, constructing synthetic data for these unobserved weeks would require complex predictive modeling, introducing significant aleatory uncertainty that could obscure the direct causal effect of the voting rules (Modeling Uncertainty).

3 Notations

Table 1: List of Notations

Symbol	Description	Symbol	Description
s	Season index	τ	Softmax temperature parameter
w, t	Week index within a season	κ	Dirichlet concentration strength
i	Contestant (celebrity) index	Ω_t	Feasible set under elimination constraint
A_t	Set of active contestants in week t	AR_t	Acceptance rate in sampling
$N_{s,w}$	TV viewership in week w of season s	$\mu_{i,t}$	Posterior mean estimate
$V_{s,w}$	Total fan votes in week w of season s	$l_{i,t}$	Lower credible bound
$V_{i,w}$	Fan votes for contestant i in week w	$r_{i,t}$	Relative credible interval width
$\alpha(s)$	Voting intensity coefficient in season s	$J_{i,s,t}$	Judge-score-based rank
α_{\min}	Minimum voting intensity	$F_{i,s,t}$	Fan-vote-based rank
α_{\max}	Maximum voting intensity	$S_{i,s,t}$	Composite rank score
α_{base}	Baseline voting intensity	τ	Kendall rank correlation coefficient
s_0, k	Logistic growth parameters	IR_t	Inversion rate
$F(w)$	Weekly heat factor	M_t	Number of rank comparisons
T_w	Total voting volume in week w	$\lambda(t x)$	Hazard function
$S_{i,w}$	Vote share of contestant i in week w	$\lambda_0(t)$	Baseline hazard
\mathbf{F}_t	Latent vote-share vector	β	Cox model coefficient
$\boldsymbol{\alpha}_t$	Dirichlet parameter vector	HR	Hazard ratio
$\alpha_{i,t}$	Dirichlet parameter of contestant i	α_s	Dynamic fan weight
$\alpha_{0,t}$	Total Dirichlet concentration	α_{final}	Fan weight in final rounds
$\pi_{i,t}$	Softmax-normalized preference score	ρ	Bonus score decay coefficient
$s_{i,t}$	PCA first component score	ω	Team score weight coefficient

4 Problem 1: Fan Votes Estimation

In *Dancing with the Stars* (DWTS), fan voting determines elimination and reflects social engagement. However, since official voting figures have never been publicly released, this core variable remains a "black box". To overcome this data limitation and accurately reconstruct the weekly voting landscape, we propose a **Two-Stage Decomposition Framework**. We break down this complex inverse problem into two observable or inferable sub-problems:

1. Predict the **Total Voting Volume**.
2. Estimate the **Voting Share Distribution**.
3. Multiply these components yields the final prediction results.

4.1 Seasonal Logistic Adopt Model : Predict the Total Voting Volume

To accurately estimate the total voting volume for each elimination week, we developed a data-driven framework that correlates viewership metrics with voting behavior. This approach calibrates historical data against verified "ground truth" anchors to account for the evolution of voting technologies and viewer engagement patterns.

4.1.1 Data Sources and Ground Truth Anchors

We utilize season-average and event-specific viewership numbers as the primary proxy for the potential voter base. To calibrate our model, we extracted official voting figures from press releases for the Season 33 and Season 34 finales. Table 2 summarizes the anchor data used to establish the baseline voting intensity.

Season	Event	Viewership ($N_{s,w}$)	Official Votes ($V_{s,w}$)	Implied Intensity (α)
Season 33	Finale	$\approx 6.36 M$	$\approx 32M$	5.03
Season 34	Finale	$\approx 9.43 M$	$\approx 72M$	7.63

Table 2: Ground Truth Anchors for Model Calibration

4.1.2 Modeling Voting Intensity (α)

We define the *Voting Intensity Coefficient*, α , as the ratio of total votes to viewership. Based on the ground truth data, we calculate a baseline coefficient for the modern era:

$$\alpha_{base} = \frac{\alpha_{33} + \alpha_{34}}{2} \approx 6.33 \quad (1)$$

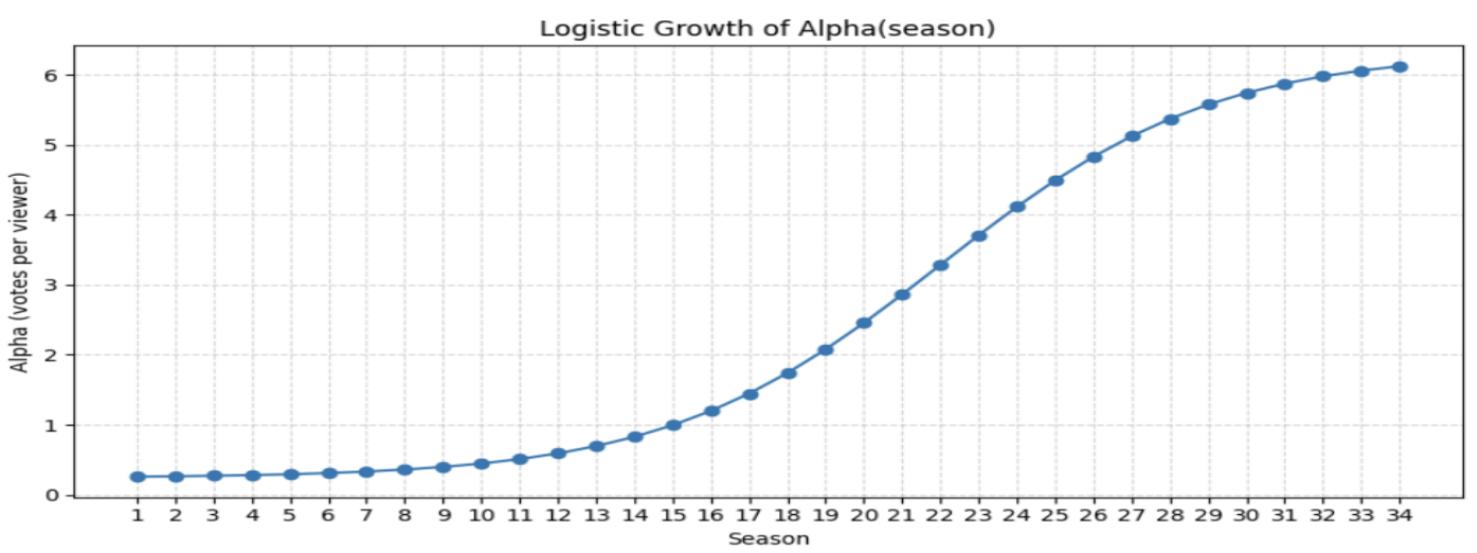
Historical analysis reveals a structural shift in voting mechanics: early seasons had high viewership but limited voting means, whereas recent seasons exhibit lower viewership but significantly higher engagement due to digital platforms allowing multiple votes per person. This trend reflects the broader shift in audience behavior in the streaming age, as noted by [1], where engagement metrics have decoupled from traditional linear viewership. Furthermore, research by [2] suggests that in such televised contests, active participation (voting) often correlates positively with the intensity of viewership engagement, even if the absolute number of passive viewers declines. We categorize this evolution into three distinct eras, as detailed in Table 3.

To capture the nonlinear adoption of digital voting technologies, we model $\alpha(s)$ using a logistic growth function anchored at $\alpha_{min} = 0.24$ and $\alpha_{max} = 6.33$:

$$\alpha(s) = \alpha_{min} + \frac{\alpha_{max} - \alpha_{min}}{1 + \exp(-k(s - s_0))} \quad (2)$$

where parameters k and s_0 control the growth rate and the inflection point of digital adoption, respectively.

Era (Seasons)	Key Context	Parameter $\alpha(s)$
Telephony Era(S1–S10)	Landlines/SMS only; 1 vote limit; fees applied. High viewers, low conversion.	0.5 – 0.8
Transition Era(S11–S27)	Online voting intro; easing limits.	Linear: $1.0 \rightarrow 4.0$
Digital Era(S28+)	Multi-channel (App/Text/Online); multiple votes allowed.	Logistic $\rightarrow [5.0, 7.0]$

Table 3: Evolution of Voting Intensity Coefficient $\alpha(s)$ Figure 3: $\alpha - s$

4.1.3 Weekly Heat Factor $F(w)$

Recognizing that viewer engagement fluctuates within a season, we introduce a Weekly Heat Factor $F(w)$ to account for the increased participation during high-stakes weeks. Table 4 presents the multipliers applied to different competition stages.

Competition Stage	Description	Factor $F(w)$
Regular Weeks	Standard elimination rounds	1.0
Semi-Finals / Finals	High-stakes episodes with peak engagement	1.2 – 1.5

Table 4: Weekly Heat Factor $F(w)$ Settings

4.1.4 Final Estimation Formula and Result

Integrating these components, the estimated total votes $V_{s,w}$ for week w of season s is calculated as:

$$V_{s,w} = \text{Round} (N_{s,w} \times \alpha(s) \times F(w)) \quad (3)$$

This formula effectively resolves the "Viewership-Vote Paradox" by explicitly modeling the inflation of vote counts in the digital era despite declining traditional TV viewership.

We visualized the estimation results in Figure 4.

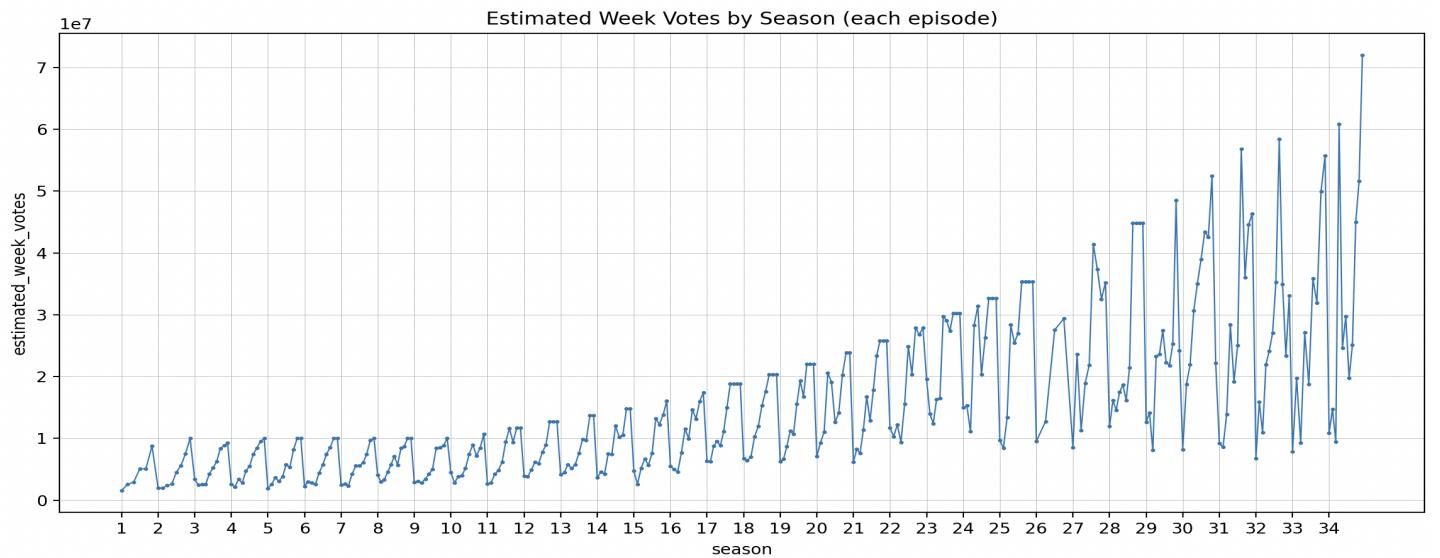


Figure 4: final result

4.2 Truncated Dirichlet Posterior model : Estimating the Voting Share Distribution

In this section we explain the design of the second model in our pipeline. Considering that the share of votes for each celebrity in each week add up to one, we adopt **Dirichlet Distribution** for a prior distribution of shares. To utilize the information hidden in the features described in the given dataset, we uses One-Hot Encoding and PCA to learn the primary information in each feature vector, and then use the first component as the determining factor that gives α , the parameter in Dirichlet Distribution. This incorporation of primary component into the parameter of Dirichlet Distribution enables the prior distribution to have a bias toward the correct direction. Then we apply the method of rejection sampling, with criterion of rejecting the ones violating the actual elimination each week, and calculate the mean of feasible points as our prediction each week.

4.2.1 Feature Engineering

In the table from the COMAP official website, we find information about the celebrity, its co-dancer, and the judge scores in each week's contest. We preprocess the data largely by implementing One-Hot encoding on pure string values except celebrity names, and replace the four judge scores each week with their mean values. There are mainly two reasons for employing One-Hot encoding:

1. One-Hot encoding has the facility of not corrupting the inter-dimensional correlation. For example, if we encode celebrity industry as 1, 2..., N then the feature has a bias that the difference between each industry is almost the same because the intervals measure the same on that dimension, which is often wrong. We assume that celebrity name won't give important information on voting, so we dropped it.
2. We can run PCA on the encoded feature vector to get the primary information the features are conveying. There are 156 unique values except celebrity name, which is just good for PCA. Celebrity name itself has 408 unique values, which is so large that it affects PCA result, so we drop it.

The main preprocessing steps are shown in Figure 5. There's one thing to clarify, though, that since either combining scores by rank or by percentage involves considering the sum of judge scores, we simplify the representation of the judge score for a celebrity in a week with the mean of the corresponding judge scores.

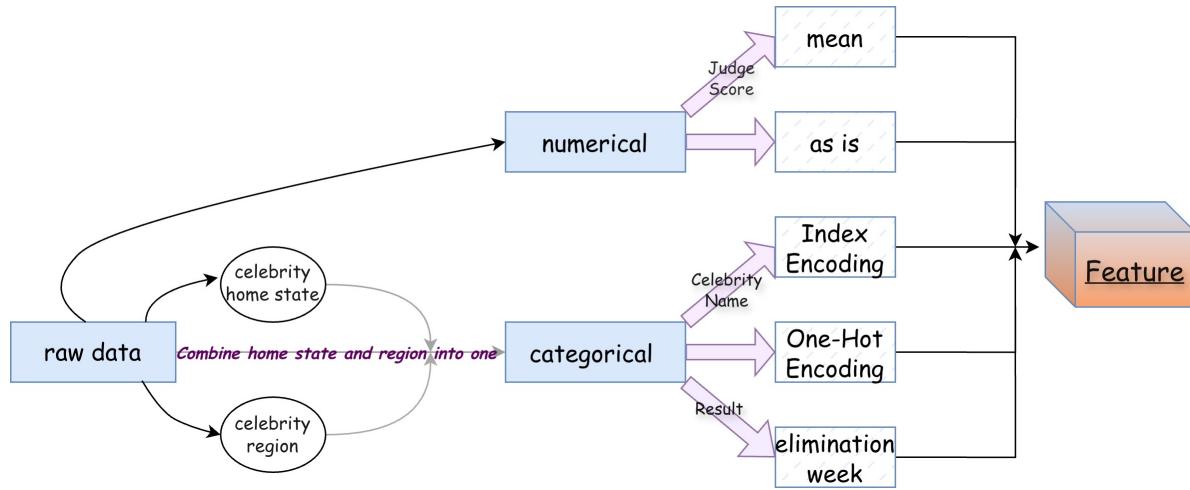


Figure 5: Flow chart of feature engineering.

4.2.2 Constructing Dirichlet Distribution

Dirichlet Distribution is known to be commonly used as prior distributions in Bayesian Statistics. It is a family of continuous multivariate probability distributions parameterized by a vector α of positive reals, denoted as:

$$\mathbf{x} = (x_1, \dots, x_K) \sim \text{Dir}(\boldsymbol{\alpha}), \quad \boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_K), \quad \alpha_i > 0,$$

with constraint on input vector:

$$\text{support: } x_i \geq 0, \quad \sum_{i=1}^K x_i = 1, \quad (4)$$

The PDF of $X \sim \text{Dir}(\alpha)$ could be formulated as:

$$p(\mathbf{x} | \boldsymbol{\alpha}) = \frac{1}{B(\boldsymbol{\alpha})} \prod_{i=1}^K x_i^{\alpha_i - 1}, \quad (5)$$

, where $B(\boldsymbol{\alpha})$ is the *Beta* function [3].

The voting events could be abstracted as each individual voter voting randomly under certain distribution independently, and this adheres to a **Multinomial Distribution**. The good news is the Dirichlet is the **conjugate prior** for the Multinomial parameters, so it fits our problem really well.

To instantiate the Dirichlet prior each week, we map features to a weekly preference share and then to $\boldsymbol{\alpha}_t$. For each (i, t) , start from the 165-dim feature matrix and standardize each column, and add an ϵ to avoid zero divider:

$$\tilde{X}_{rj} = \frac{X_{rj} - \mu_j}{\sigma_j + \epsilon} \quad (6)$$

Apply PCA on \tilde{X} and take the first component score as $s_{i,t}$, then normalize within each (season, week) to eliminate cross-week difference:

$$s'_{i,t} = \frac{s_{i,t} - \mu_t}{\sigma_t + \epsilon} \quad (7)$$

Convert $\{s'_{i,t}\}_{i \in A_t}$ into a simplex preference by temperature-softmax:

$$\pi_{i,t} = \frac{\exp(s'_{i,t}/\tau)}{\sum_{k \in A_t} \exp(s'_{k,t}/\tau)} \quad (8)$$

and finally set the Dirichlet parameters with concentration strength κ :

$$\alpha_{i,t} = 1 + \kappa \pi_{i,t} \quad (9)$$

$$\alpha_{0,t} = \sum_{i \in A_t} \alpha_{i,t} = |A_t| + \kappa. \quad (10)$$

4.2.3 Rejection Sampling

For each season-week t , we perform rejection sampling to obtain samples from a truncated Dirichlet posterior that is consistent with the observed elimination outcome. Let

$$\mathbf{F}_t \sim \text{Dir}(\boldsymbol{\alpha}_t) \quad (11)$$

denote the latent vote-share vector on the simplex. Define the feasible set Ω_t as the subset of the simplex where the contestant(s) with the lowest share(s) are exactly the true eliminated contestant(s) in week t . We then repeat:

$$\mathbf{F}_t^{(m)} \sim \text{Dir}(\boldsymbol{\alpha}_t) \quad (12)$$

$$\text{accept } \mathbf{F}_t^{(m)} \iff \mathbf{F}_t^{(m)} \in \Omega_t \quad (13)$$

otherwise reject and resample. The procedure continues until M accepted samples are obtained (or a preset trial cap is reached), yielding an empirical approximation of the truncated posterior $\text{Dir}(\boldsymbol{\alpha}_t) \mid (\mathbf{F}_t \in \Omega_t)$. The mean of the accepted samples each week is used as the estimated value of the share vector, which contains different celebrities' shares in that week.

4.2.4 Multiply these components yields the final prediction results

$$V_{i,w} = T_w \times S_{i,w} \quad (14)$$

where T_w represents the total voting volume derived from viewership metrics, and $S_{i,w}$ denotes the normalized voting share ($\sum S_{i,w} = 1$) estimated from contestant performance.

4.3 Consistency Analysis

Since our *final* share estimates are posterior means under a truncated Dirichlet posterior conditioned on observed eliminations, the estimations we make is 100% consistent with the eliminations each week. But still we can measure the consistency of the underlying prior distribution, with **accept rate** defined by

$$AR_i = \frac{\#\text{accepted}_i}{\#\text{trials}_i} \quad (15)$$

The results are given in Fig 6, which implies that our feature-determined α contributes to a good preference.

4.4 Certainty Analysis

For the measurement of certainty, we adopt the statistic of **Relative Credible Interval Width** to show how much our estimation of fan votes can be trusted. The ReCI is given as

$$r_{i,t} = \frac{\mu_{i,t} - l_{i,t}}{\hat{f}_{i,t} + \epsilon} \quad (16)$$

, which is the normalized measurement of the scale of possible values. From the distribution illustrated in Fig 6 we can derive the fact that there are differences in certainty across (week, contestant) pairs. Smaller ReCIs means narrower set of possible estimations, and greater certainty. Our result shows that mostly the ReCIs stay less than 3.3, with the mean being approximately 2.6. This is not a small value, we confess, but as fan votes are unobserved and we only have loose weekly constraint on elimination, this result is natural.

5 Problem 2 : Comparative Analysis of Voting Methods

5.1 Data Processing

To ensure the robustness of our comparative analysis, we performed rigorous data preprocessing. The raw dataset was filtered to retain only "elimination weeks," excluding non-elimination rounds .

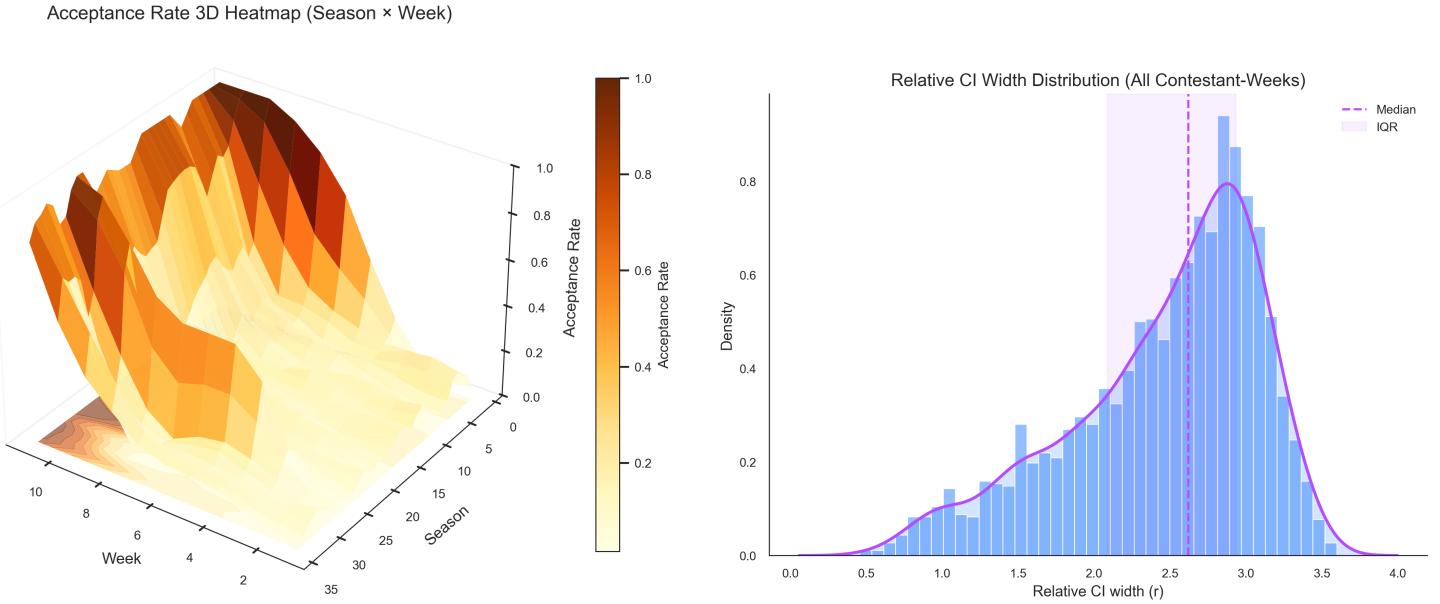


Figure 6: Heat map of accept rate week-wise and distribution of ReCI widths.

5.2 Experimental Assumption: Week-Local Counterfactual Framework

To rigorously evaluate the impact of voting methodologies, we adopt a **Week-Local Counterfactual** approach rather than a "Full-Season Replay." This decision is grounded in a fundamental identifiability constraint:

- **Missing Data Problem:** In a full-season replay, if a contestant is counterfactually "saved" from elimination in Week t , their performance scores and fan votes for Week $t + 1$ do not exist in the historical dataset.
- **Modeling Uncertainty:** Constructing synthetic data for these unobserved weeks would require complex predictive modeling (e.g., estimating hypothetical judge scores), introducing significant aleatory uncertainty that could obscure the direct causal effect of the voting rules.

Assumption: We assume that the "fairness" of a voting system can be assessed by its immediate decision-making behavior in each independent elimination event.

5.3 Quantifying Methodological Divergence

To rigorously measure the difference between the two ranking systems, we employed Kendall's Rank Correlation Coefficient (τ) [4][5]. This statistic evaluates the similarity between two orderings of the same set of objects.

The coefficient is defined as:

$$\tau = \frac{2}{N(N-1)} \sum_{i < j} \text{sgn}(x_i - x_j) \text{sgn}(y_i - y_j) \quad (17)$$

where N is the number of active contestants in the current week, and (x_i, y_i) are the ranks of the i -th contestant under the two methods.

Table 5: Interpretation of Kendall's Rank Correlation Coefficient (τ)

Value	Interpretation
$\tau = 1$	Perfect agreement (identical rankings)
$\tau = 0$	No correlation (random relationship)
$\tau = -1$	Perfect disagreement (inverse rankings)

Since τ can be unintuitive for interpreting the magnitude of conflict, we introduced the **Inversion Rate** (IR_t), which quantifies the percentage of contestant pairs whose relative ordering is flipped between the two methods:

$$IR_t = \frac{1 - \tau}{2} \quad (18)$$

An IR_t of 0 implies perfect consistency, while higher values indicate greater divergence.

To obtain a season-level metric, we calculated the weighted average IR_{season} , where weights M_t correspond to the number of pairwise comparisons ($M_t = \frac{N(N-1)}{2}$) in week t :

$$IR_{season} = \frac{\sum_t M_t IR_t}{\sum_t M_t} \quad (19)$$

We established the following thresholds to categorize the significance of the divergence:

Table 6: Thresholds for Methodological Divergence

Inversion Rate (IR)	Category	Interpretation
$IR < 0.05$	Negligible	Methods yield practically identical results.
$0.05 \leq IR < 0.15$	Significant	Noticeable shifts in mid-tier rankings.
$IR \geq 0.15$	Substantial	Fundamental disagreement on outcomes.

Result Analysis: Our calculations reveal a season-average IR of approximately 11.3%. According to our thresholds (Table 6), this indicates a **Significant Difference** between the Percentage and Ranking methods.

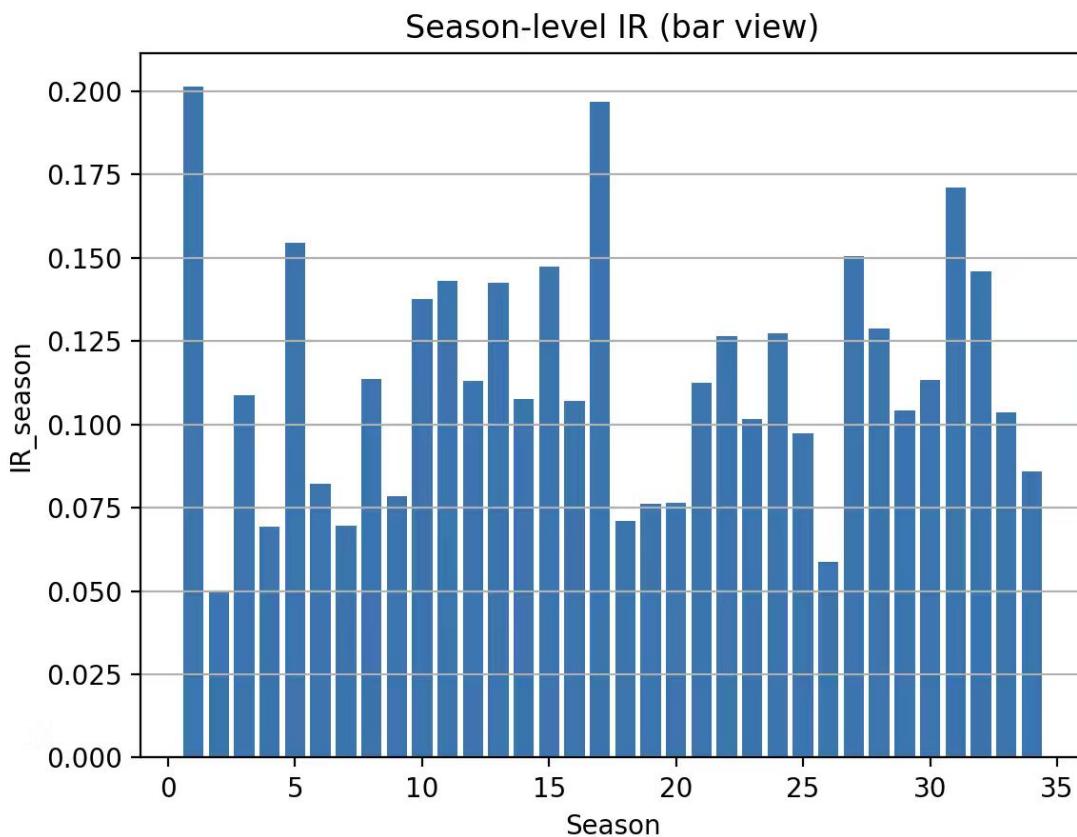


Figure 7: Visualizing the divergence between voting approaches across seasons.

5.4 Fan Influence Analysis: Which Approach Favors the Audience?

The core appeal of reality competitions lies in the agency of the fanbase. To determine which method empowers fans more, we developed the **Fan Dominance Index (FDI)**. The FDI measures the relative alignment of the final ranking (R_{method}) with the fan vote ranking (R_F) versus the judges' ranking (R_J):

$$FDI_{method} = \frac{\tau(R_F, R_{method}) - \tau(R_J, R_{method})}{2} \quad (20)$$

A positive *FDI* indicates that the method aligns more closely with fan preferences, while a negative value suggests a bias towards the judges. We aggregated this metric to the season level using the same weighting scheme M_t :

$$FDI_{season} = \frac{\sum_t M_t FDI_t}{\sum_t M_t} \quad (21)$$

Findings: As illustrated in Figure 8, the Percentage Method consistently yields higher positive FDI values compared to the Ranking Method. This conclusively demonstrates that **the Percentage Method favors fans**, allowing their voting magnitude to override judges' scores more effectively than the rank-based approach.

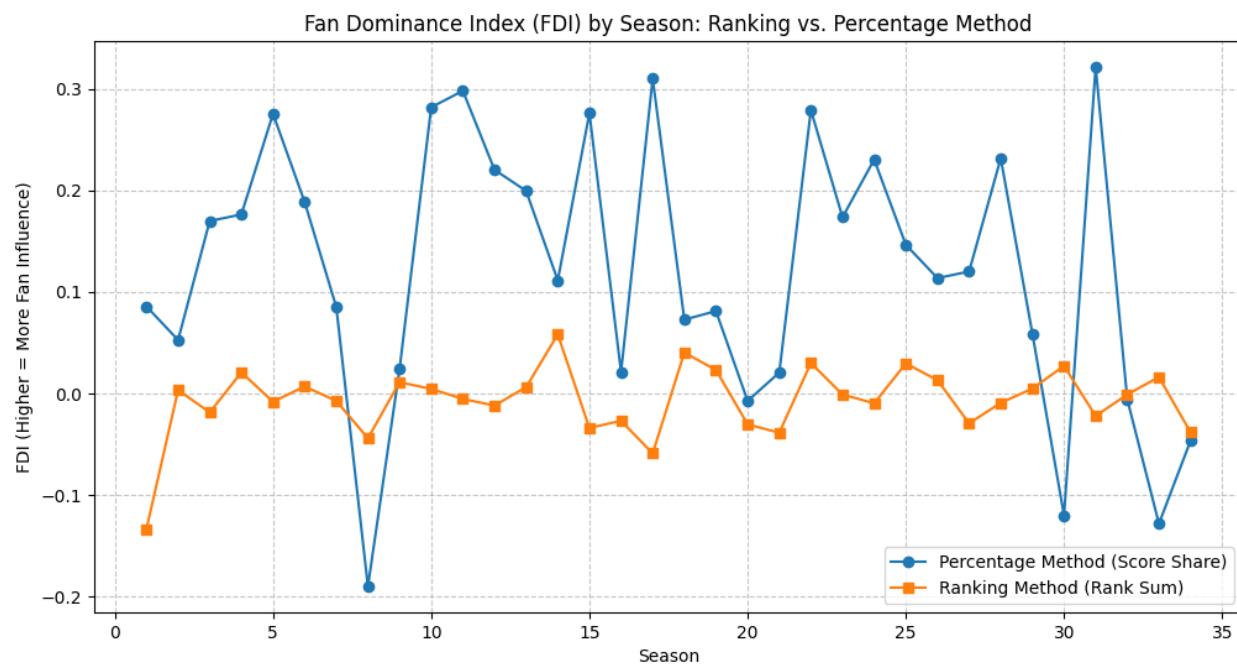


Figure 8: Fan Dominance Index (FDI) by Season.

5.5 Case Studies: Divergence in Outcomes

To illustrate the practical impact of these methods, we analyzed four specific contestants (ID 11, 35, 126, 305) who exemplify the tension between judge scores and fan popularity: **Jerry Rice (Season 2)** **Billy Ray Cyrus (Season 4)** **Bristol Palin (Season 11)** and **Bobby Bones (Season 27)**.

Detailed Case Analysis

We examined four controversial cases to test methodological impacts.

- **Jerry Rice (S2) & Billy Ray Cyrus (S4):** Both received consistently low judge scores but high fan support. The **Ranking Method** punished their technical deficits early (e.g., placing Cyrus in the

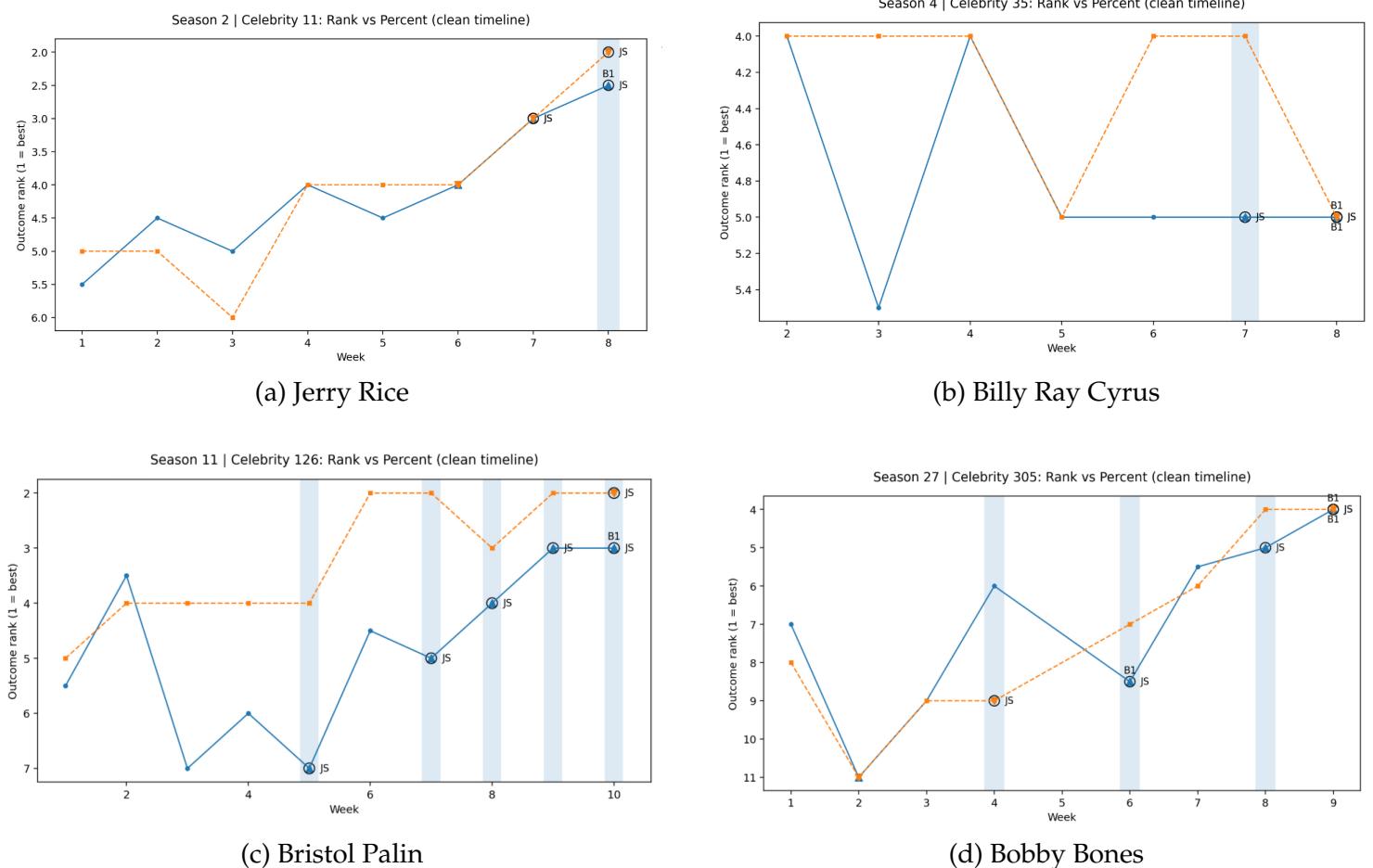


Figure 9: rank change of four cases

Blue solid line → Rank method outcome rank
 Orange dashed line → Percent method outcome rank
 Blue upward-pointing triangle → Rank bottom 2
 Orange downward-pointing triangle → Percent bottom 2

Bottom 2 by Week 7), whereas the **Percentage Method** allowed their fan vote magnitude to override judge scores, delaying their risk of elimination.

- **Bristol Palin (S11):** A prime example of "fan insulation." Under the Ranking Method, she fell into the Bottom 2 frequently (Weeks 5, 7, 8, 9). Conversely, the Percentage Method shielded her almost entirely until Week 10, confirming it favors polarized contestants with strong fan bases.
- **Bobby Bones (S27):** The Ranking Method would have eliminated him as early as Week 6. The Percentage Method sustained him until Week 9, highlighting its tendency to favor "populist" candidates over technical merit.

Impact of the "Judges' Save"

Our simulation identifies the "Judges' Save" as a **meritocratic corrective**. While the Percentage Method often protects popular low-scorers from direct elimination (Bottom 1), it frequently fails to keep them out of the "Bottom 2." Once in this risk zone, the "Judges' Save" restores power to the experts, who typically eliminate the contestant with the lower technical rank, effectively neutralizing the "fan shield" effect.

Conclusion: The Percentage Method amplifies fan influence, extending the survival of popular but technically weaker contestants. The Ranking Method enforces stricter technical standards. The Judges' Save serves as a critical fail-safe to balance popularity with skill.

5.6 Recommendations for Future Seasons

To determine the optimal voting methodology for future seasons, we evaluated the options based on two conflicting objectives: **Competitive Fairness** (meritocracy) and **Commercial Viability** (viewership/engagement).

First, regarding fairness, the Percentage Method risks overshadowing talent with popularity, evidenced by a high Fan Dominance Index (peaking at **0.32 in Season 31**). The Ranking Method corrects this by capping extreme vote disparities, ensuring technical merit remains central.

Next, concerning viewership, our **11.3% Inversion Rate** confirms that the Ranking Method stabilizes results but still allows fans to alter outcomes in significant cases. This maintains suspense and engagement without compromising competitive integrity.

Conclusion: We recommend the **Ranking Method** to mitigate popularity landslides while preserving meaningful fan agency.

5.7 Evaluation of the "Judges' Save" Proposal

We strongly endorse the addition of the "Judges' Save" mechanism for the bottom two couples. We summarize three reasons:

Conflict Maximization: Viewership thrives on conflict. The current system provides conflict between fans and judges. Adding a "Judges' Save" introduces a new layer of **survival conflict** between the bottom two couples directly.

Heightened Uncertainty: Our simulations reveal that while final outcomes often converge, the **intermediate safety margins** vary significantly. For example, controversial contestants (like **ID 305**) would face repeated "survival duels" rather than coasting on fan votes, keeping the audience engaged until the final second.

Narrative Drama: The mechanism prevents anticlimactic eliminations. Our case study analysis shows that **all of 4 cases in part 5.5** experienced drastically different survival trajectories under this rule, allowing for more complex "redemption arcs" and preventing premature exits of high-potential talent.

6 Problem 3: Impact of celebrity characteristics and various professional dancers on performance

We preprocessed data to the task needs, and constructed a bipartite model. In this task we borrow the preprocessed data in Task 1, but due to the objective of impact analysis we selected only celebrity's home, age, industry, and ballroom_partner. We apply **PCA** to each categorical feature to reduce dimensionality to 1 respectively. Then we concatenated them into a 4D feature vector. Then a **Cox Proportional Hazards Model** is employed to evaluate the overall impact on performance, while for judge-fan difference we use a **Stacked Linear Regression** to evaluate the difference. This is essentially two linear regression on both judge scores and fan votes but will make it clearer to see the difference between the two coefficients.

6.1 Data Processing: feature-wise PCA

There are three categorical features out of the four concerning this task, each encoded in One-Hot manner, so it's hard to directly use the 165-d feature vector for analysis. For one thing, high dimensionality leads to under-fitting considering our dataset scale; for another, features with dimensionality higher than 2 will hamper feature-wise impact analysis.

Thus we applied PCA to each three categorical features to reduce dimensionality to 1, and concatenated them into a 4-d feature vector, which is ready for fitting and analysis.

6.2 The Cox PH Model: How much is the impact?

The **Cox Proportional Hazards Model**[6] is called a survival model, measuring the probability that a hazard occurs at time t . Like any other survival model, it has two parts: the underlying baseline hazard function, denoted as $\lambda_0(t)$, describing how hazard risk changes over time at baseline level; the effect pa-

rameters, describing how hazard risk behave in response to covariates. The definition of it could be given as

$$\lambda(t | x) = \lambda_0(t) \exp(\beta^\top x). \quad (22)$$

, as long as the *proportional hazards condition* holds, which is also an important property:

$$\frac{\lambda(t | x_1)}{\lambda(t | x_2)} = \exp(\beta^\top (x_1 - x_2)), \quad \forall t \geq 0 \quad (23)$$

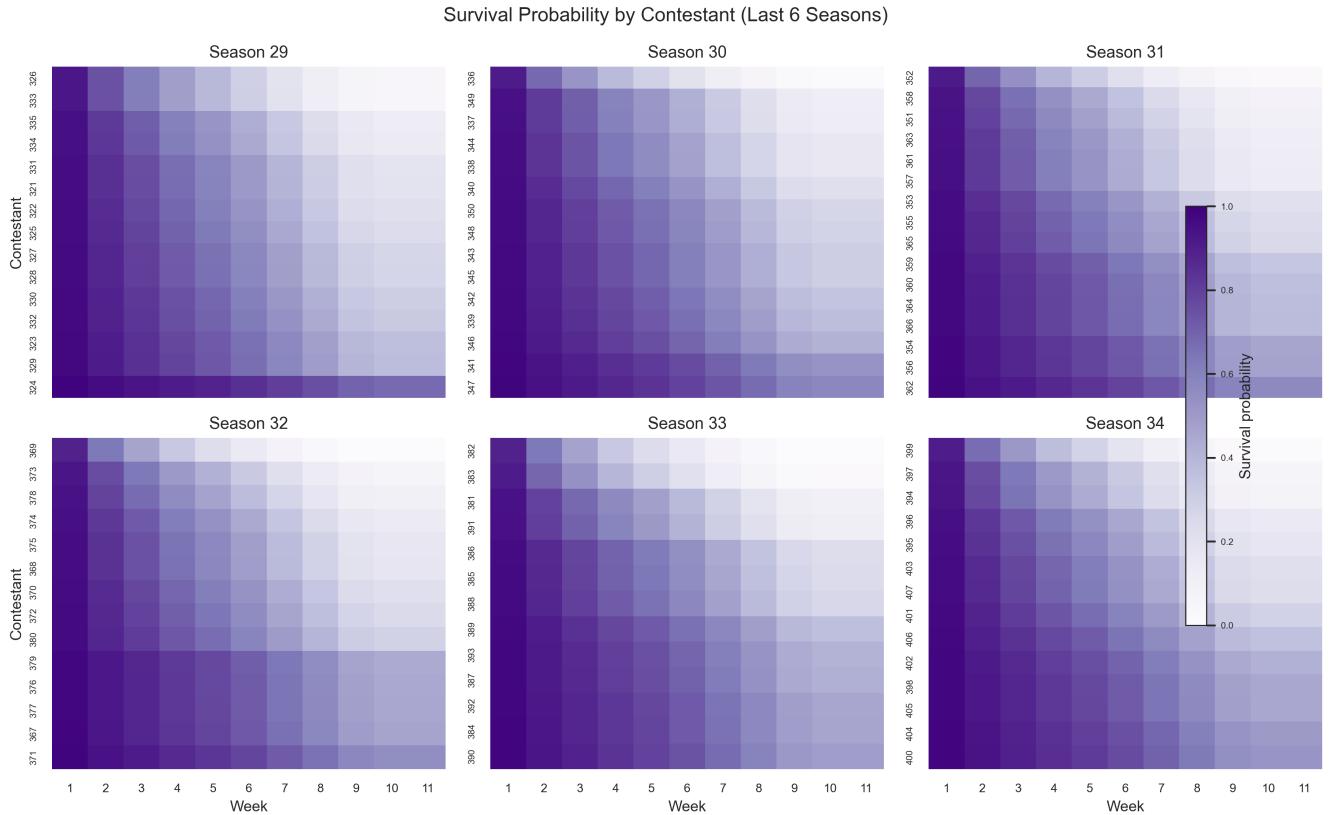


Figure 10: Survival Rate of Last 6 Seasons Across Contestants and Week

By our assumption discussed before, we can apply this model to the analysis of performance, with the covariate being our preprocessed 4D feature vector and the hazard being elimination. Eliminated in week t can be equivalently treated as hazard happening in week t , and the elimination week can represent the performance of a celebrity.

In our specific implementation, we define the event time T_i as the week in which celebrity i was eliminated. Observations where the contestant reaches the end of the season without elimination are treated as right-censored ($\delta_i = 0$).

To estimate the coefficients β without specifying the functional form of the baseline hazard $\lambda_0(t)$, we utilize the concept of **Partial Likelihood**. A specific challenge in this dataset is the presence of “tied” event times, where multiple contestants are eliminated in the same week. To address this, we employ **Efron’s approximation**, which provides a more accurate estimate of the risk set in the presence of ties compared to the standard Breslow approximation. The log-partial likelihood with Efron’s correction is given by:

$$\ell(\beta) = \sum_j \left(\sum_{i \in D_j} \beta^\top x_i - \sum_{k=0}^{d_j-1} \ln \left(\sum_{\ell \in R_j} e^{\beta^\top x_\ell} - \frac{k}{d_j} \sum_{\ell \in D_j} e^{\beta^\top x_\ell} \right) \right) \quad (24)$$

where D_j is the set of individuals failing at time t_j (with size d_j), and R_j is the risk set at time t_j .

From the fitted coefficients, we derive the **Hazard Ratio (HR)** for each feature, defined as $HR_k = \exp(\beta_k)$. **HR’s value reflects how much impact each feature has on performance**. For example, from

table 7 we can infer that celebrity's name matters the most concerning performance.

Table 7: Hazard ratios (HR), transposed

	Z_industry	Z_home	Z_pro	age_z
HR	1.158827	0.871386	1.053537	1.741168

Finally, to visualize the performance trajectory, we reconstruct the survival function $S(t|x)$, which represents the probability of a contestant *surviving* past week t . We estimate the baseline cumulative hazard $H_0(t)$ using the **Breslow estimator**[7]:

$$\hat{H}_0(t) = \sum_{t_j \leq t} \frac{d_j}{\sum_{\ell \in R_j} \exp(\hat{\beta}^\top x_\ell)} \quad (25)$$

The survival probability for a specific contestant with feature vector x is then computed as $\hat{S}(t|x) = \exp(-\hat{H}_0(t)e^{\hat{\beta}^\top x})$. This allows us to generate heatmaps (Fig10) comparing the survival prospects of contestants based purely on their static characteristics (Industry, Home, Pro Partner, and Age).

6.3 Different impact on judge scores and fan votes

To rigorously quantify the divergence between professional evaluation and public popularity, we employ a **Stacked Linear Regression** framework. While running two separate regressions (one for judges, one for fans) would yield individual coefficient estimates, it would not provide a direct statistical test for the *difference* between those coefficients. Stacking the data allows us to model this difference as an interaction term within a unified system.

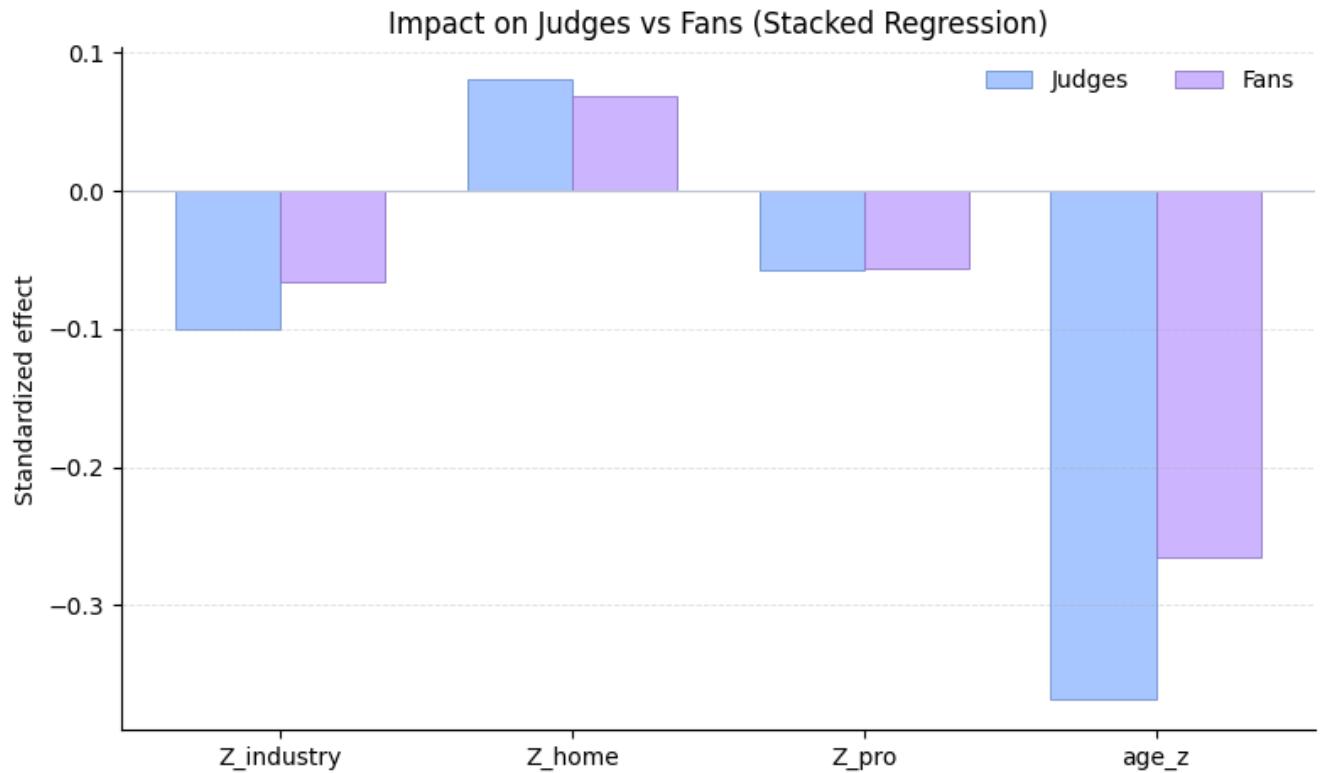


Figure 11: Stacked regression result on feature-wise impact on both judge scores and fan votes

First, we harmonize the response variables to a common scale. The judge scores, denoted as Y_J , are standardized to zero mean and unit variance. For fan votes, provided as a voting share $S \in (0, 1)$, we first

apply the logit transformation to map the bounded probability space to the real line:

$$Y_F^* = \ln\left(\frac{S}{1-S}\right) \quad (26)$$

The transformed fan metric Y_F^* is subsequently standardized to create the final response Y_F . We then construct a stacked dataset of size $2N$, introducing an indicator variable D_i where $D_i = 0$ if the observation corresponds to a judge score and $D_i = 1$ if it corresponds to a fan vote.

The resulting model is defined as:

$$E[Y] = \alpha + \gamma D + \beta^\top X + D(\delta^\top X) \quad (27)$$

where α is the global intercept, γ captures the baseline shift between the two metrics, and X is the 4D feature vector derived from PCA.

In this formulation, the vector β represents the effect of characteristics on **Judges' scores**, while δ captures the **differential effect** for Fans. Specifically, the total effect of a feature on fan voting is given by $\beta + \delta$. Table 8 shows that the impacts on judge scores and fan votes are indeed different. δ on the age dimension even reaches 0.1.

Table 8: Differential Impact of Features on Fan Votes vs. Judge Scores ($\delta = \beta_{fan} - \beta_{judge}$)

Feature	Estimate (δ)	S.E.	p-value
Industry Score	0.034	(0.021)	0.109
Home State Score	-0.012	(0.021)	0.555
Pro Partner Score	0.000	(0.021)	0.994
Age (Standardized)	0.102***	(0.021)	< 0.001

Significance codes: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$

7 Problem 4: Optimization of the Scoring and Elimination System

To address the limitations of the current system—specifically the "Viewership-Vote Paradox" and the occasional elimination of high-performing contestants due to low popularity—we propose a new comprehensive framework. This framework is structurally divided into two primary mechanisms:

1. An optimized **Elimination Mechanism**, which integrates a **Dynamic Weighted Rating (DWR)** system to balance professional judgment and audience engagement dynamically.
2. A **Mid-Season Resurrection Mechanism**, designed to provide a second chance for high-quality contestants who were prematurely eliminated.

7.1 Notation and Definitions

We define the following variables to formalize our proposed model:

- s, t, i : Indices representing the season, week, and contestant, respectively.
- T_{SStart} : The start week of the season S .
- T_{SEnd} : The final week of the season S .
- $J_{i,s,t}$: The rank of contestant i based on judge scores in week t of season s .
- $F_{i,s,t}$: The rank of contestant i based on fan votes in week t of season s .
- V_s : The total volume of votes cast in season s .
- α_s : The dynamic weight coefficient for fan voting in season s .
- $S_{i,s,t}$: The final composite score for contestant i in week t of season s .

7.2 Elimination Policy: Rank-Based with Judge Save

We retain the **Bottom-Two Judge Save** mechanism. The two contestants with the lowest composite scores $S_{i,s,t}$ face elimination. The judges then vote to save one, adding a final layer of protection for skilled dancers who may have had a single bad week in terms of popularity.

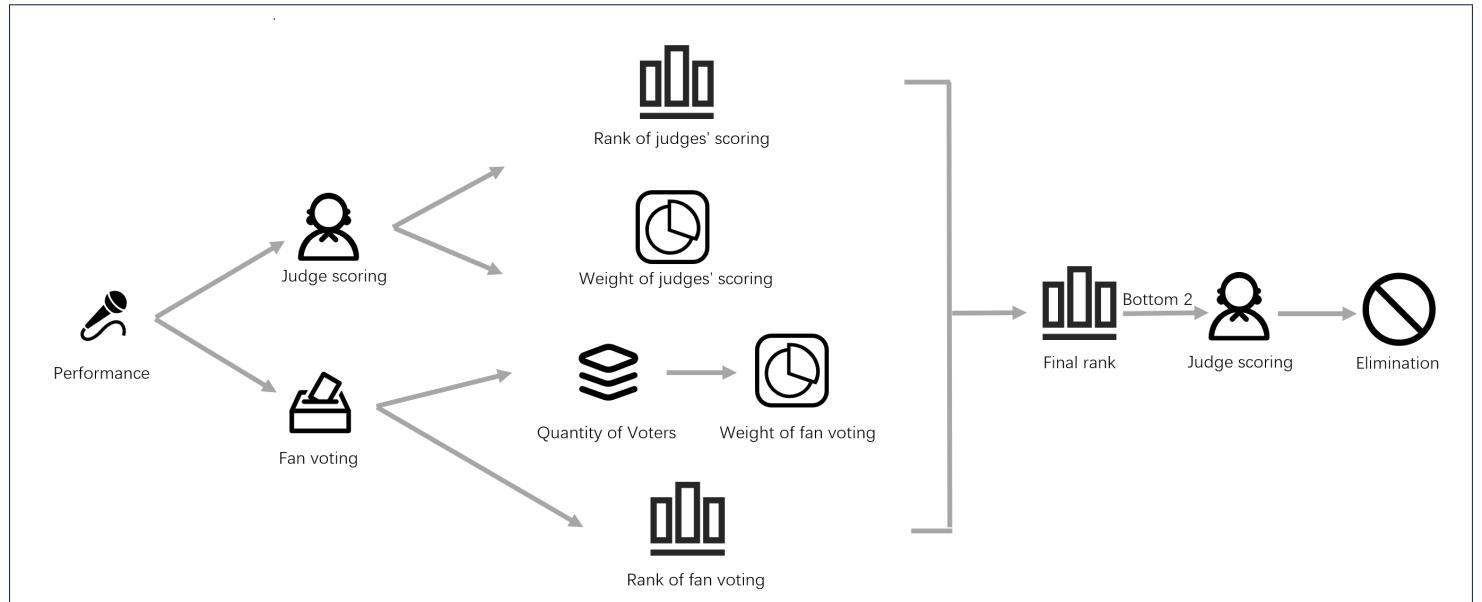


Figure 12: elimination mechanism

7.3 Dynamic Weighted Rating (DWR) Mechanism

We propose the **Dynamic Weighted Rating (DWR)** system, which adjusts the influence of fan votes based on the total voting engagement of the season and the stage of the competition.

7.3.1 Dynamic Weight Calculation (α_s)

For regular weeks ($t < T_{SEnd} - 1$), the weight of fan votes, α_s , is not constant. Instead, it is a function of the normalized total voting volume \tilde{V}_s . We use a sigmoid-based function to model this relationship, ensuring that seasons with higher voter engagement grant more power to the audience, while still bounding the influence to prevent popularity contests from completely overshadowing skill.

$$\alpha_s = \alpha_{\min} + (\alpha_{\max} - \alpha_{\min}) \cdot \sigma(a + b \cdot \tilde{V}_s) \quad (28)$$

where:

- $\alpha_{\min} = 0.40$ and $\alpha_{\max} = 0.65$ represent the lower and upper bounds of fan influence.
- $\sigma(\cdot)$ is the standard sigmoid function.
- $b > 0$ ensures that higher voting volume leads to higher fan weight.
- a, b are parameters that can be fitted to historical data or set to robust constants ($a = 0, b = 1$ for baseline).

The composite score for regular weeks is calculated as a weighted sum of ranks:

$$S_{i,s,t} = (1 - \alpha_s) \cdot J_{i,s,t} + \alpha_s \cdot F_{i,s,t}, \quad \text{for } t < T_{s-1} \quad (29)$$

7.3.2 Championship Integrity Protection

For the Finals ($t \geq T_s - 1$), we explicitly prioritize professional judgment and objectivity to ensure the winner is technically deserving. We fix the fan weight to a lower value α_{final} :

$$S_{i,s,t} = (1 - \alpha_{\text{final}}) \cdot J_{i,s,t} + \alpha_{\text{final}} \cdot F_{i,s,t} \quad (30)$$

where $\alpha_{\text{final}} = 0.35$. This grants judges 65% of the decision power, preventing a "popularity winner" who lacks dance proficiency.

7.3.3 Handling Bonus and Team Scores

To further refine the scoring, we adjust how external points (Bonus and Team Dance) are integrated to avoid disproportionate impacts.

1. Bonus Points: Let $B_{i,s,t}$ be the raw bonus points awarded. We introduce a decay coefficient $\rho \in [0.3, 0.6]$ to scale its impact:

$$P_{i,s,t}^{\text{bonus}} = \rho \cdot B_{i,s,t} \quad (31)$$

We recommend conducting sensitivity analysis on ρ to determine the optimal value that rewards excellence without distorting the rank-based baseline.

2. Team Dance Scores: Team dances introduce external dependencies where an individual's score is tied to others. Let $T_{i,s,t}$ be the team score assigned to the individual. We down-weight this component using $\omega \in [0.3, 0.7]$:

$$P_{i,s,t}^{\text{team}} = \omega \cdot T_{i,s,t} \quad (32)$$

7.4 Resurrection Mechanism: The "Mid-Season Wildcard"

7.4.1 Motivation

Historical data reveals instances where high-performing contestants were eliminated prematurely due to low fan support. Notable examples include:

- **Sam Champion (Season 31, Week 4):** Judge Rank #1, Fan Rank #10 → Eliminated.
- **Sabrina Bryan (Season 15, Week 6):** Judge Rank #1, Fan Rank #8 → Eliminated.

These cases highlight a flaw where extreme disparities between skill and popularity lead to the loss of top-tier talent. To mitigate this, we introduce a resurrection mechanism.

7.4.2 Mechanism Design

The resurrection pool $\mathcal{P}_{\text{pool}}$ is formed from contestants eliminated in the interval $[T_{SStart}, T_{SEnd} - 1]$.

Selection Process:

1. **Meritocracy (Judge's Priority):** The top 2 eliminated contestants based on cumulative judge scores (S_{judge}) enter the pool.
2. **Popularity (Fan's Choice):** From the remaining eliminated contestants, the top 2 based on cumulative fan votes (S_{fan}) enter the pool.

Final Decision: The judges select one contestant from $\mathcal{P}_{\text{pool}}$ to return to the competition, ensuring the resurrected candidate contributes to the competitive quality of the finals.

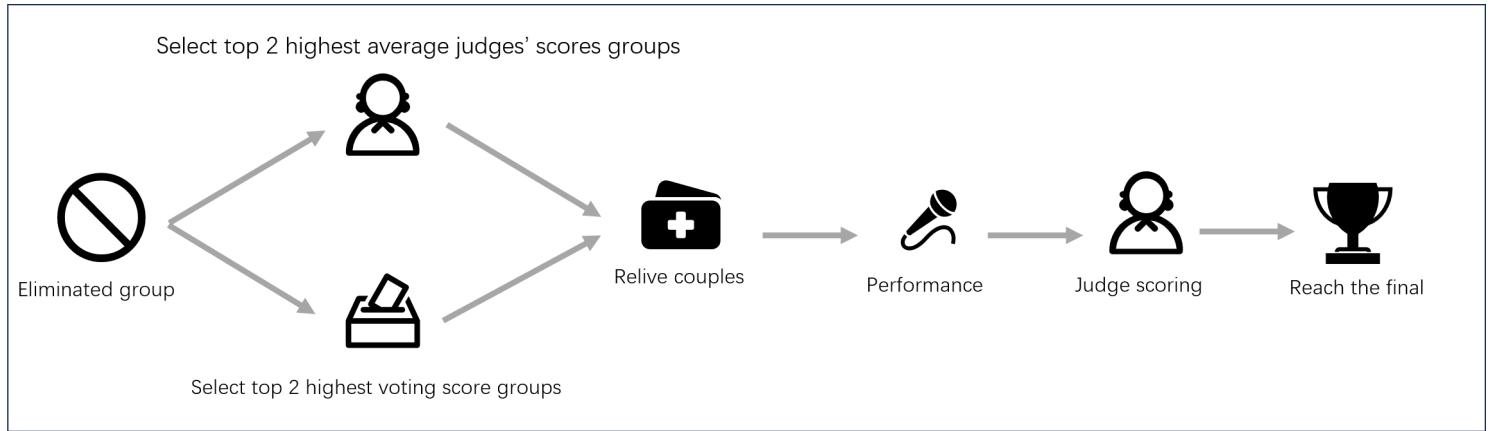


Figure 13: resurrection mechanism

8 Sensitivity Analysis

The key sensitive parameters are from the **Truncated Dirichlet Posterior Model**'s hyper parameters. These parameters are chosen not based on facts but arbitrarily. Therefore, we performed a sensitivity analysis on the fan-share estimator by sweeping the softmax temperature τ and Dirichlet strength κ . Specifically, we ran the full rejection-sampling pipeline on real DWTS data for $\tau \in \{0.3, 0.5, 0.8, 1.2\}$ and $\kappa \in \{5, 10, 20\}$. For each grid point we computed weekly entropies and correlations to the baseline share estimates.

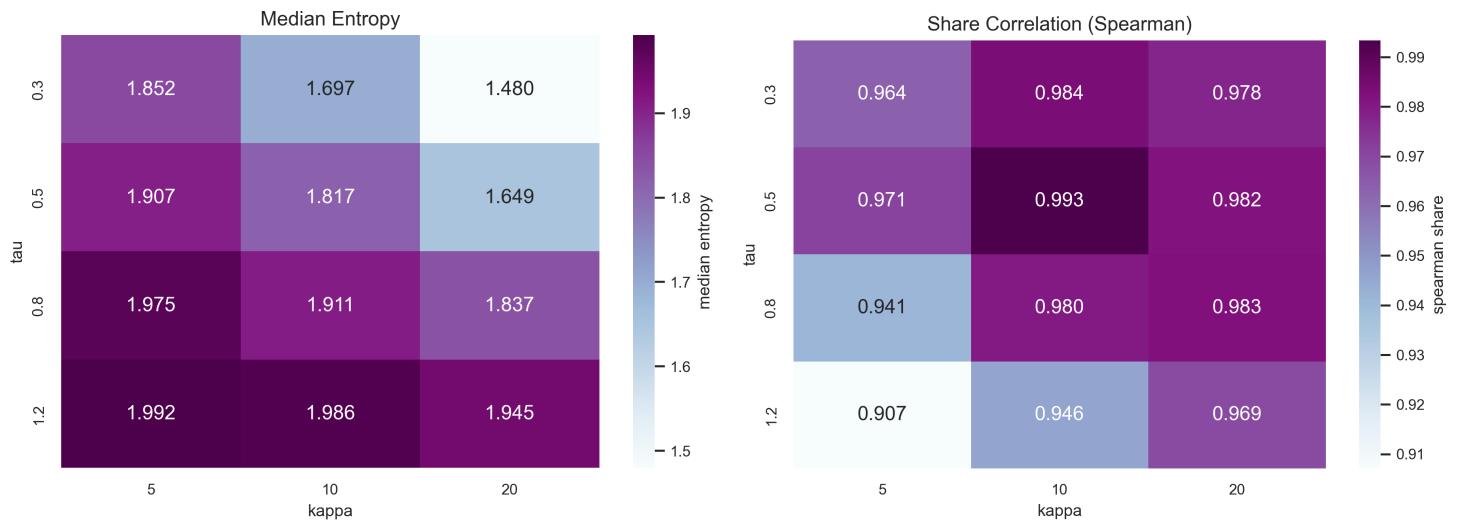


Figure 14: Grid: Entropy and Spearman Correlation Heatmap.

Across the grid, entropy varies smoothly, and share correlations to baseline remain high. This indicates the inference is **robust** to reasonable changes in τ and κ , and the reported fan-share estimates are not artifacts of a particular prior setting.

As a more concrete and clear example, we display a randomly chosen celebrity's statistics in Fig 9. The fan votes share are all above 0.51 and under 0.68, which is a small range, and indicates that our model gives stable predictions with varying hyper-parameters.

Table 9: share_mean of Kelly Monaco indexed by (τ, κ) in season 1 week 6.

τ	$\kappa = 5$	$\kappa = 10$	$\kappa = 20$
0.3	0.518109	0.544658	0.583403
0.5	0.559899	0.607100	0.674982
0.8	0.570811	0.610656	0.648590
1.2	0.539026	0.571089	0.618689

9 Strengths and weaknesses

In this section we discuss some of the strengths and weaknesses of our models.

9.1 Strengths

Our fan votes prediction model has much strength on predictions.

- In the estimation of fan votes, the hardest thing is that there are literally no real fan votes given to us. Our model ensures feasibility by first estimating the total votes each week based on fragmented data we found online and a regression model.
- The prior Dirichlet distribution has no preference, and simply sampling from this prior distribution will yield estimations that are merely consistent, with a large variance. In our model we processed the feature vector we extracted from the MCM dataset and transformed this into a reasonable preference of the Dirichlet Distribution. In this way the distribution now has a preference based on our information at hand, which makes the estimations more feasible and with smaller variance.

As for the pro-dancer's and celebrity characteristics' impact on performance and judge-fan difference, our model has the strength of evaluating impact on a feature-wise scale, clearly displaying how each feature contributed to the impact of performance, or judge scores and fan votes.

9.2 Weaknesses

The biggest weakness of our models is in the characteristic impact analysis, i.e., in the third problem. Before we fitted the **Cox PH Model** and the **Stacked Linear Regression Model**, we applied PCA to reduce each One-Hot encoded feature's dimension to 1. This might cause great information losses but we must choose between mixing the features together at the cost of low interpretability and discriminating different features at the cost of losses of much information. Also our model cannot tell, say, whether being Athlete is better or being Actor/Actress because of PCA.

10 Memorandum

To: Dancing with the Stars Production Team

From: Team #2608504

Subject: Data-Driven Recommendations on Combining Judges' Scores and Fan Votes

Date: February 2, 2026

Dear Members of the Production Team,

This memorandum summarizes our quantitative analysis of the voting and elimination mechanisms used in *Dancing with the Stars* (DWTS), based on comprehensive data from Seasons 1–34. Our objective is to assess how different methods of combining judges' scores and fan votes affect competitive fairness, outcome stability, and audience engagement, and to provide actionable recommendations for future seasons.

Key Findings on Fan Voting Behavior. Because official fan vote counts are not publicly released, we reconstructed weekly fan voting using a two-stage framework. First, total weekly vote volume was estimated from viewership data using a calibrated voting intensity coefficient, anchored by verified finale vote counts from Seasons 33 and 34. This analysis indicates that modern DWTS seasons operate in a high-engagement digital regime, with an average of approximately **6.33 votes per viewer**, reflecting the widespread adoption of multi-channel and repeat voting.

Second, weekly fan vote shares among contestants were inferred by estimating relative preference proportions and multiplying these shares by the estimated total vote volume. The share allocation follows a probabilistic structure equivalent to modeling contestant vote proportions with a Dirichlet-type prior, which is statistically natural for multi-candidate voting processes. By conditioning these share estimates on observed eliminations, the reconstructed fan votes are fully consistent with historical outcomes while still exhibiting meaningful variation across contestants and weeks.

Impact of Vote-Combination Rules. We conducted a systematic comparison of the two historical combination rules used by DWTS: the *rank-based method* and the *percentage-based method*. Using a week-local counterfactual framework, we found that the two approaches produce a season-average inversion rate of approximately **11.3%**, indicating a statistically significant divergence in weekly rankings and elimination risk.

Further analysis using the Fan Dominance Index (FDI) shows that the **percentage-based method consistently aligns final outcomes more closely with fan vote rankings than with judges' rankings**. In contrast, the rank-based method caps the influence of extreme fan vote disparities, preserving the relative importance of technical performance. Case studies—including Jerry Rice (Season 2), Bristol Palin (Season 11), and Bobby Bones (Season 27)—demonstrate that the percentage-based method can substantially prolong the survival of highly popular but lower-scoring contestants, whereas the rank-based method enforces stricter technical accountability.

Role of the Judges' Save Mechanism. Our simulations indicate that the **Judges' Save** serves as an effective meritocratic safeguard. While fan-driven systems may protect popular contestants from immediate elimination, they frequently place such contestants in the bottom two. In these cases, allowing judges to select which couple is eliminated consistently mitigates extreme popularity bias, stabilizes outcomes, and prevents technically strong performers from premature exit. This mechanism also increases dramatic tension without undermining competitive credibility.

Impact of Professional Dancers and Celebrity Characteristics. Using the cleaned multi-season dataset, we constructed a bipartite representation of celebrities and professional dancers and reduced high-dimensional categorical characteristics via feature-wise PCA. A Cox Proportional Hazards model was then employed to quantify how static celebrity attributes and partner effects influence elimination risk, identifying characteristics associated with longer competitive survival. In parallel, a stacked regression framework was used to measure how these characteristics differentially affect judges' scores and fan votes, enabling the identification of contestants with strong performance potential but asymmetric audience or judge response.

Recommendations for Future Seasons. Based on these findings, we recommend that DWTS:

- **Adopt the rank-based method** as the primary mechanism for combining judges' scores and fan votes, as it balances audience participation with technical merit and limits excessive popularity-driven distortions.
- **Retain the Judges' Save mechanism**, particularly during mid-season elimination rounds, to correct extreme mismatches between fan popularity and dance quality.
- **Avoid fully percentage-based aggregation** in early and middle weeks, where it is most likely to amplify fan vote landslides and reduce perceived fairness.

These recommendations preserve meaningful fan influence while ensuring that competitive outcomes remain credible, skill-oriented, and narratively compelling.

Sincerely yours,
Team #2608504

References

- [1] M. Barrett, C. Shao, and H. Mantas, "Revisiting the impact of structural factors on television audience behavior in the streaming age," *Journal of Broadcasting & Electronic Media*, vol. 67, no. 1, pp. 1–20, 2023.
- [2] J. A. Amegashie, "American idol: Should it be a singing contest or a popularity contest?" *Journal of Cultural Economics*, vol. 33, no. 4, pp. 265–277, 2009.
- [3] S. Kotz, N. Balakrishnan, and N. L. Johnson, *Continuous Multivariate Distributions, Volume 1: Models and Applications*, 2nd ed. New York: Wiley, 2000, ISBN: 978-0-471-18387-7.
- [4] M. G. Kendall, "A new measure of rank correlation," *Biometrika*, vol. 30, no. 1–2, pp. 81–89, 1938. DOI: 10.1093/biomet/30.1-2.81.
- [5] W. H. Kruskal, "Ordinal measures of association," *Journal of the American Statistical Association*, vol. 53, no. 284, pp. 814–861, 1958. DOI: 10.2307/2281954.
- [6] N. E. Breslow, "Analysis of survival data under the proportional hazards model," *International Statistical Review / Revue Internationale de Statistique*, vol. 43, no. 1, pp. 45–57, 1975. DOI: 10.2307/1402659.
- [7] D. Y. Lin, "On the breslow estimator," *Lifetime Data Analysis*, vol. 13, pp. 471–480, 2007. DOI: 10.1007/s10985-007-9048-y.

Appendices

Appendix A Estimated fan votes of last season

Celebrity	Wk	Votes	Celebrity	Wk	Votes
Alix Earle	2	1028301	Jen Affleck	3	501020
Alix Earle	3	671954	Jen Affleck	4	275563
Alix Earle	4	4564596	Jen Affleck	6	348268
Alix Earle	6	2408785	Jen Affleck	7	271032
Alix Earle	7	1673606	Jen Affleck	8	254543
Alix Earle	8	2327074	Jen Affleck	9	266033
Alix Earle	9	4575990	Jen Affleck	10	168571
Alix Earle	10	5512765	Jen Affleck	11	237078
Alix Earle	11	7588463	Jordan Chiles	2	600006
Andy Richter	2	704243	Jordan Chiles	3	366090
Andy Richter	3	153459	Jordan Chiles	4	241903
Andy Richter	4	337888	Jordan Chiles	6	187291
Andy Richter	6	105715	Jordan Chiles	7	148405
Baron Davis	2	538940	Jordan Chiles	8	106346
Baron Davis	3	243859	Jordan Chiles	9	262040
Baron Davis	4	103021	Jordan Chiles	10	200725
Baron Davis	6	107905	Jordan Chiles	11	195152
Baron Davis	7	96856	Lauren Jauregui	2	1141261
Baron Davis	8	108664	Lauren Jauregui	3	141118
Baron Davis	9	166807	Lauren Jauregui	4	688850
Baron Davis	10	85150	Lauren Jauregui	6	777660
Baron Davis	11	103609	Lauren Jauregui	7	664535
Corey Feldman	2	79559	Lauren Jauregui	8	576654
Danielle Fishel	2	717883	Lauren Jauregui	9	712776
Danielle Fishel	3	494573	Lauren Jauregui	10	492765
Danielle Fishel	4	3251625	Lauren Jauregui	11	590784
Danielle Fishel	6	1705467	Robert Irwin	2	814677
Danielle Fishel	7	1274298	Robert Irwin	3	540971
Danielle Fishel	8	1527530	Robert Irwin	4	357951
Danielle Fishel	9	2869470	Robert Irwin	6	239914
Danielle Fishel	10	3323486	Robert Irwin	7	196416
Danielle Fishel	11	4144806	Robert Irwin	8	215559
Dylan Efron	2	604025	Robert Irwin	9	185090
Dylan Efron	3	311858	Robert Irwin	10	112458
Dylan Efron	4	154427	Robert Irwin	11	157195
Dylan Efron	6	301306	Scott Hoying	2	714410
Dylan Efron	7	139516	Scott Hoying	3	457488
Dylan Efron	8	196458	Scott Hoying	4	3236305
Elaine Hendrix	2	2050481	Scott Hoying	6	182240
Elaine Hendrix	3	345358	Scott Hoying	7	240916
Elaine Hendrix	4	596519	Scott Hoying	8	323691
Elaine Hendrix	6	1671237	Scott Hoying	9	525805
Elaine Hendrix	7	1262360	Scott Hoying	10	421469
Elaine Hendrix	8	1691915	Scott Hoying	11	250413
Elaine Hendrix	9	3129056	Whitney Leavitt	2	1117421
Elaine Hendrix	10	4028508	Whitney Leavitt	3	747809
Elaine Hendrix	11	4979697	Whitney Leavitt	4	119420
Hilaria Baldwin	2	698782	Whitney Leavitt	6	684481
Hilaria Baldwin	3	482647	Whitney Leavitt	7	465584
Hilaria Baldwin	4	310657	Whitney Leavitt	8	422349
Jen Affleck	2	741641	Whitney Leavitt	9	532432
			Whitney Leavitt	10	731011

Table 10: Estimated fan votes by (celebrity, week).

Report on the Use of AI

1. Google Gemini (February 2, 2026 version, Gemini-3-Pro-Preview)

Query 1: Figure Height Adjustment

Query: Can you help me keep the width of the figure and shorten the height?

Output: To shorten the height of the figure while maintaining the width at 0.5\textwidth , you can explicitly set the height parameter in the `\includegraphics` command.

```
\begin{figure}[H]
\centering
% Set height explicitly. Adjust '4cm' to your desired height.
% Note: This will distort the image aspect ratio.
\includegraphics[width=0.5\textwidth, height=4cm]{3.1.5.result.png}
\caption{final result}
\label{fig:final_result}
\end{figure}
```

Alternatively, to crop the image without distortion:

```
% trim = <left> <bottom> <right> <top>
\includegraphics[width=0.5\textwidth, trim=0 2cm 0 2cm, clip]{3.1.5.result.png}
```

Query 2: Header Adjustment

Query: How to adjust the header in LaTeX?

Output: Use the `fancyhdr` package.

```
\usepackage{fancyhdr}
\pagestyle{fancy}
\fancyhf{} % Clear existing headers and footers
\fancyhead[L]{My Document Title}
\fancyhead[R]{\thepage}
\renewcommand{\headrulewidth}{0.4pt}
```

2. OpenAI ChatGPT

Query 3: Handling Boolean Values in Pandas

Query: Can boolean values in pandas data frame be properly handled by packages like sklearn and seaborn?

Output: The assistant explained that for MCM-style competition data, a person's name is a high-cardinality categorical identifier. Encoding it directly is usually ineffective or leaky. The recommended approach is to treat names as entities and extract history-based skill features.

Key recommendations:

1. Treat names as entities and extract history-based skill features (e.g., `person_prev_mean_score`).
2. Use hashing or target encoding for categorical features if necessary.
3. Encode pair relationships explicitly.

Query 4: Parallel Figures in LaTeX

Query: In latex, how to arrange two figures parallel horizontally. I mean, one left and one right so that they fill up the horizontal space.

Output: Use `subfigures` (recommended) so the two images sit left and right.

```
\usepackage{subcaption}
\begin{figure}[htbp]
\centering
\begin{subfigure}[t]{0.49\textwidth}
\centering
\includegraphics[width=\linewidth]{fig_left.png}
\caption{Left caption}
\label{fig:left}
\end{subfigure}\hfill
\begin{subfigure}[t]{0.49\textwidth}
\centering
\includegraphics[width=\linewidth]{fig_right.png}
\caption{Right caption}
\label{fig:right}
\end{subfigure}
\caption{Overall caption for both.}
\label{fig:two_side_by_side}
\end{figure}
```

Query 5: Beautiful Figures in Python

Query: How to draw beautiful figures in Python? I'm working on MCM 2026 Problem C.

Output: To make MCM figures look publication-grade, focus on consistent styling, clear encodings, clean layout, and correct export.

```
import matplotlib as mpl
import matplotlib.pyplot as plt

def set_mcm_style():
    mpl.rcParams.update({
        "figure.dpi": 120,
        "savefig.dpi": 300,
        "figure.figsize": (7.2, 3.2),
        "font.size": 11,
        "axes.spines.top": False,
        "axes.spines.right": False,
        "axes.grid": True,
        "grid.alpha": 0.25,
    })
set_mcm_style()
```
