Problem Chosen	2026 MCM/ICM Summary Sheet	Team Control Number
C		2608125

Reconstructing Hidden Fan Votes and Stress-Testing Voting Rules in Dancing with the Stars: A Stable-ABC Bayesian Pipeline and FairVote Backtest

Summary

Televised dance competitions combine observed judges' scores with undisclosed fan votes, making the elimination process only partially observable. Using multi-season data from *Dancing with the Stars* (DWTS), we develop a unified modeling pipeline to reconstruct latent fan vote shares, quantify uncertainty, compare alternative vote-combination rules, explain performance/popularity drivers, and propose a tunable new voting mechanism.

For Task 1, we infer week-level fan vote shares on the probability simplex via a symmetric Dirichlet prior and a Stable-ABC (Approximate Bayesian Computation) procedure that enforces the season-specific historical elimination rule. Posterior summaries yield vote-share means and uncertainties, with week-level diagnostics (acceptance rate and average posterior standard deviation) indicating inference difficulty. A posterior predictive check achieves perfect agreement with observed eliminations across 229 elimination weeks when posterior means are replayed through the historical rule.

For Task 2, we perform counterfactual replays by holding inferred fan shares fixed while switching the aggregation rule among `Rank`, `Percent`, and `Bottom2+JudgeSave`. We quantify consistency, flip events, and contestant-level survival-week shifts, showing that `Percent` provides the most stable and interpretable magnitude-based aggregation, while judge-save mechanisms can induce qualitatively larger outcome shifts even when `Rank` and `Percent` agree.

For Task 3, we fit ridge-regularized linear models with cross-validation to explain three targets: `WeeksSurvived`, `AvgJudgeTotal`, and `AvgFanShare`. Grouped coefficient magnitudes highlight that the professional partner is the strongest signal for judge outcomes and remains influential for fan support, with moderate contributions from industry and geographic attributes; sensitivity analyses confirm robustness to regularization and resampling.

For Task 4, we propose `FairVote`, a two-parameter rule that explicitly trades off historical consistency, "close-call" excitement rate, and fan-protection behavior. A grid backtest with bootstrap confidence intervals and leave-one-season-out checks identifies an operating point that yields moderate consistency while maintaining a controlled intervention rate, enabling stakeholders to choose policy priorities transparently. Overall, our framework provides an auditable approach for reconstructing hidden preferences, diagnosing rule sensitivity, and supporting evidence-based voting-rule design.

Keywords: Bayesian inference, ABC, fan votes, voting rules, fairness

Contents

1 Introduction

1.1 Background

Televised dance competitions combine performance and popularity under a weekly elimination system. lemphDancing with the Stars (DWTS) is a representative example, where outcomes depend on judges' scores and fan votes. Judges' scores are observed and standardized, but fan vote totals are not disclosed and remain proprietary. Thus, the elimination process is only partially observable: eliminations are known, while the underlying vote distribution is hidden. This motivates mathematical modeling. By reconstructing plausible vote shares consistent with observed outcomes, we can evaluate how different combination rules affect fairness, predictability, and controversy. We also examine whether contestant attributes and professional partners influence success beyond dancing performance.

1.2 Literature Review

Voting-based competitions have been examined through social choice theory and fairness analysis, especially in how rankings or normalized scores are aggregated into final outcomes. When important variables are not directly observed, Bayesian inference and simulation-based methods such as Approximate Bayesian Computation (ABC) provide practical tools to reconstruct latent preferences with uncertainty.

1.3 The Description of the Problem

We study DWTS eliminations across multiple seasons, where weekly outcomes depend on judges' scores and undisclosed fan votes. The dataset provides judge scores, contestant attributes, season/week identifiers, and the eliminated couple each week.

Our objectives are:

1. **Fan Vote Inference:** Estimate weekly fan vote shares consistent with observed eliminations and quantify uncertainty.
2. **Voting Rule Comparison:** Compare rank-based and percentage-based systems, and test a bottom-two judge decision mechanism.
3. **Feature Impact Analysis:** Evaluate how professional partners and celebrity characteristics relate to scores, popularity, and survival.
4. **Rule Design:** Propose and justify a fairer voting procedure supported by quantitative evidence.

1.4 Our Work

We propose a unified pipeline combining Bayesian reconstruction, rule-based simulation, and interpretable predictive modeling.

Task 1: Fan vote reconstruction. We infer latent fan vote shares for each active couple using a Dirichlet prior and an ABC-style procedure that generates plausible vote vectors consistent with observed eliminations.

Task 2: Rule comparison. Using inferred fan preferences, we replay historical weeks under multiple aggregation rules and summarize agreement rates, flip cases, and disagreement patterns.

Task 3: Feature impact. We use regularized regression to separate performance-driven effects from popularity-driven effects and identify influential contestant and partner attributes.

Task 4: New voting system proposal. Based on empirical findings, we propose and backtest an alternative voting scheme that aims to improve fairness while preserving competition excitement.

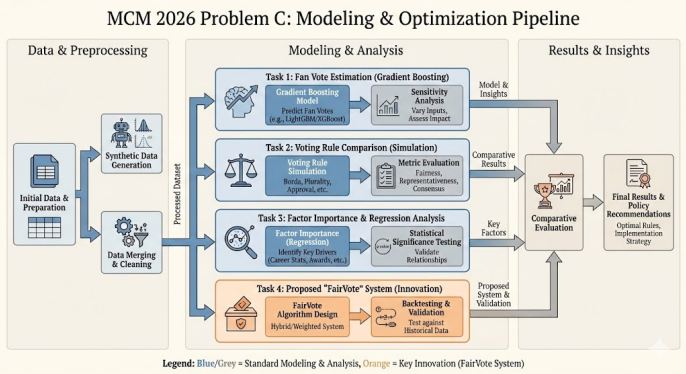


Figure 1: Modeling Pipeline Overview

2 Assumptions

Assumption 1: (Data correctness & structure) The Excel file `2026_MCM_Problem_C_Data.csv` is treated as the single source of truth. We assume judge score columns (when present) are recorded correctly, and the `results` field correctly indicates the eliminated week in the format "Eliminated Week {w}".

Assumption 2: (Active roster definition) A couple is considered *active* in a given week if they have not been eliminated before that week. Weeks where the eliminated couple cannot be uniquely identified from the data are excluded from inference and downstream comparisons.

Assumption 3: (Judge score aggregation) A couple's weekly judge score is computed by aggregating all available judge scores for that week. Missing entries are ignored, and invalid values (e.g., non-positive placeholders) are treated as absent rather than informative.

Assumption 4: (Fan votes are unobserved) Fan votes are not directly observed in the dataset. We treat fan support as a latent (hidden) quantity that influences eliminations together with judges' scores.

Assumption 5: (Season-dependent rule regime) We assume the show follows different elimination rule regimes across seasons: an early-season ranking-based regime, a mid-season percentage-based regime, and a later-season "bottom-two with judges' save" regime. The regime is assumed constant within each season.

Assumption 6: (Rule implementation is a faithful proxy) When applying elimination rules, we use an operational implementation that matches the intended mechanics:

- `Percent`: combine normalized judges' contribution with fan support and eliminate the lowest combined score.
- `Rank`: combine judges and fan support via rank-based aggregation and eliminate the worst combined rank.
- `Bottom2+JudgeSave`: identify a bottom-two set by combined risk and assume judges save the couple with the higher judges' score.

We assume this implementation is sufficiently faithful for modeling and comparison purposes, including how ties are handled.

Assumption 7: (Week-wise independence) Fan support is inferred independently week by week, conditional on that week's active roster and judges' scores. We do not explicitly model temporal dynamics such as popularity momentum across weeks.

3 Notations

The core symbols and their definitions used in this study are summarized in Table 1, providing an overview of the key parameters and their related meanings.

Table 1: Notations used in this model

Symbol	Description
s	Season index (e.g., $s = 1, 2, \dots$)
w	Week index within a season ($w = 1, 2, \dots, W$)
W	Maximum week number detected from the input columns
i	Index of an active couple in week w (among $n_{s,w}$ active couples)
$n_{s,w}$	Number of active couples with valid judge scores in season s, week w
$\text{CoupleID}_{s,w,i}$	Identifier of couple i in season s, week w (Celebrity_Partner)
$J_{s,w,i}$	Total judge score of couple i in week w (sum of available positive judge scores)
$\mathbf{J}_{s,w}$	Vector of judge totals for all active couples in (s, w): $\mathbf{J}_{s,w} = (J_{s,w,1}, \dots, J_{s,w,n_{s,w}})$
$e_{s,w}$	Index of the *actual* eliminated couple in week w (ground truth from the dataset)
$\mathbf{v}_{s,w}$	Fan vote share vector in (s, w): $\mathbf{v}_{s,w} = (v_{s,w,1}, \dots, v_{s,w,n_{s,w}})$, $v_{s,w,i} \geq 0$, $\sum_i v_{s,w,i} = 1$
α	Dirichlet prior concentration parameter for fan shares (code: `DIRICHLET_ALPHA`)
$\text{Dirichlet}(\alpha\mathbf{1})$	Prior distribution used to draw candidate fan share vectors $\mathbf{v}_{s,w}$
$\mathcal{R}$	Voting rule / method: $\mathcal{R} \in \{\text{Rank}, \text{Percent}, \text{Bottom2+JudgeSave}\}$
$\mathcal{R}_{\text{like}}(s)$	Likelihood rule used for inference in season s (piecewise by era: Rank / Percent / Bottom2+JudgeSave)
$\widehat{e}_{s,w}(\mathcal{R})$	Predicted eliminated couple under rule $\mathcal{R}$ given $(\mathbf{J}_{s,w}, \mathbf{v}_{s,w})$
$\text{rank}(\cdot)$	Descending rank operator (best value has rank 1; ties receive averaged ranks)
N_{draw}	Number of prior draws used in the accepted phase (reported as `DrawsUsed`)
N_{acc}	Number of accepted (or resampled) posterior samples (code target: `MIN_ACCEPT`; reported as `AcceptedSamples`)

Continued on next page

Table 1 (continued)

Symbol	Description
$N_{\min}$	Minimum accepted samples required for stable posterior summary (code: `MIN_ACCEPT`)
N_{early}	Early-stop accepted threshold (code: `EARLY_STOP_ACCEPT`)
$\rho_{s,w}$	Acceptance rate in the used phase: $\rho_{s,w} = N_{\text{acc}}/N_{\text{draw}}$ (reported as `AcceptanceRate`)
$\ell_{s,w}$	Fallback level of inference: $\ell_{s,w} \in \{0, 1, 2\}$ (exact / bottom-2 membership / soft resampling)
$\mu_{s,w,i}$	Posterior mean of fan share for couple i in (s, w) (reported as `VoteMean`)
$\sigma_{s,w,i}$	Posterior standard deviation of fan share for couple i in (s, w) (reported as `VoteStd`)
$\overline{\sigma}_{s,w}$	Average posterior uncertainty in week (s, w): mean of $\{\sigma_{s,w,i}\}_{i=1}^{n_{s,w}}$ (reported as `AvgVoteUncertainty`)
$\mathbb{I}[\cdot]$	Indicator function (1 if condition true, else 0) used in match/flip calculations
$\text{Flip}(\mathcal{R}_a, \mathcal{R}_b)$	Whether two methods disagree on predicted elimination in a week (e.g., Rank vs Percent)
$E_s(\mathcal{R}, \text{CoupleID})$	Predicted elimination week of a couple under method $\mathcal{R}$ in season s (default $W + 1$ if never eliminated)
ΔE	Benefit / harm measure as elimination-week difference between methods (e.g., $E_s(\text{Percent}) - E_s(\text{Rank})$)

4 Models

4.1 Modeling and Solving of Task 1: Bayesian Fan-Vote Inference

4.1.1 Problem Analysis

For each season s and week w, suppose there are $n_{s,w}$ active couples. Let $J_{s,w,i}$ denote the (observed) total judge score for couple i in week (s, w), and let $\mathbf{v}_{s,w} = (v_{s,w,1}, \ldots, v_{s,w,n_{s,w}})$ denote the (unobserved) fan vote share, where $\mathbf{v}_{s,w}$ lies on the probability simplex:

$$v_{s,w,i} \geq 0, \quad \sum_{i=1}^{n_{s,w}} v_{s,w,i} = 1.$$

The data provide the *actual eliminated couple* $e_{s,w}$ each week (when elimination occurs), but do not provide the fan vote share $\mathbf{v}_{s,w}$. The key difficulty is that the shows elimination is a deterministic outcome of a *voting rule* combining judge scores and fan votes; hence we infer $\mathbf{v}_{s,w}$ from the observed elimination event.

4.1.2 Model Preparation

Voting-rule eras. Historical rules are season-dependent and implemented as a piecewise function:

$$\mathcal{R}_{\text{like}}(s) = \begin{cases} \texttt{Rank}, & s \leq 2, \\ \texttt{Percent}, & 3 \leq s \leq 27, \\ \texttt{Bottom2+JudgeSave}, & s \geq 28. \end{cases}$$

For each week we compute judge totals $J_{s,w,i}$ from the available judge columns (`week{w}_judge{j}_score`, summing positive scores), and collect them as $\mathbf{J}_{s,w} = (J_{s,w,1}, \ldots, J_{s,w,n_{s,w}})$.

Notation for rule outcome. For each rule $\mathcal{R} \in \mathcal{R}$, define a deterministic map

$$g_{\mathcal{R}}(\mathbf{J}_{s,w}, \mathbf{v}_{s,w}) \in \{1, \ldots, n_{s,w}\}$$

that returns the eliminated index under rule $\mathcal{R}$ with judge vector $\mathbf{J}_{s,w}$ and fan shares $\mathbf{v}_{s,w}$. This encapsulates the three implemented mechanisms:

- `Percent`: normalize judges $\tilde{J}_{s,w,i} = J_{s,w,i} / \sum_{j=1}^{n_{s,w}} J_{s,w,j}$ and eliminate $\arg\min_i(\tilde{J}_{s,w,i} + v_{s,w,i})$.
- `Rank`: rank judges and fans separately in descending order (best rank = 1), sum ranks, and eliminate the worst (largest rank-sum).
- `Bottom2+JudgeSave`: compute bottom-2 by rank-sum, then eliminate the one with lower judge score among that bottom-2.

Bayesian prior. We impose a symmetric Dirichlet prior on fan shares [2]:

$$\mathbf{v}_{s,w} \sim \text{Dirichlet}(\alpha \mathbf{1}), \quad \alpha > 0,$$

where small α encourages sparse/peaked vote distributions, matching the fact that a few couples often dominate votes.

4.1.3 Model Construction

ABC likelihood and the intractability. The observation is the eliminated index $e_{s,w} = g_{\mathcal{R}_{\text{like}}(s)}(\mathbf{J}_{s,w}, \mathbf{v}_{s,w})$. A natural (but degenerate) likelihood is an indicator:

$$\mathcal{L}(\mathbf{v}_{s,w} \mid e_{s,w}, \mathbf{J}_{s,w}) = \mathbb{I}\left[g_{\mathcal{R}_{\text{like}}(s)}(\mathbf{J}_{s,w}, \mathbf{v}_{s,w}) = e_{s,w}\right],$$

which makes the exact posterior hard to compute analytically. We therefore use Approximate Bayesian Computation (ABC) with rejection / relaxed acceptance [1, 3].

Stable three-phase ABC inference. For each season-week, we generate samples from the prior and accept them based on how well the simulated elimination matches the observed elimination. The algorithm is stabilized by (i) a minimum accepted sample target and (ii) progressively relaxed acceptance:

1. **Phase 0 (Exact ABC).** Draw $\mathbf{v}_{s,w}^{(m)} \sim \text{Dirichlet}(\alpha \mathbf{1})$ and accept if $g_{\mathcal{R}_{\text{like}}(s)}(\mathbf{J}_{s,w}, \mathbf{v}_{s,w}^{(m)}) = e_{s,w}$. Denote the number of draws actually used by N_{draw} and the number of accepted (or resampled) posterior samples by N_{acc}. Stop early once at least N_{early} samples are accepted, and require at least $N_{\min}$ accepted samples to form posterior summaries, with a hard draw cap $D_{\max}$.
2. **Phase 1 (Bottom-2 membership relaxation).** If Phase 0 yields fewer than $N_{\min}$ accepts, we relax the event: accept if the actual eliminated index belongs to the rule-defined bottom-2 risk set. This keeps inference informative while avoiding collapse when exact matching is rare.

3. **Phase 2 (Soft importance resampling).** If Phase 1 is still insufficient, we draw a large pool $\{\mathbf{v}_{s,w}^{(m)}\}_{m=1}^{M}$ and assign weights

$$w_m \propto \exp(-T \cdot \mathrm{gap}(\mathbf{v}_{s,w}^{(m)})),$$

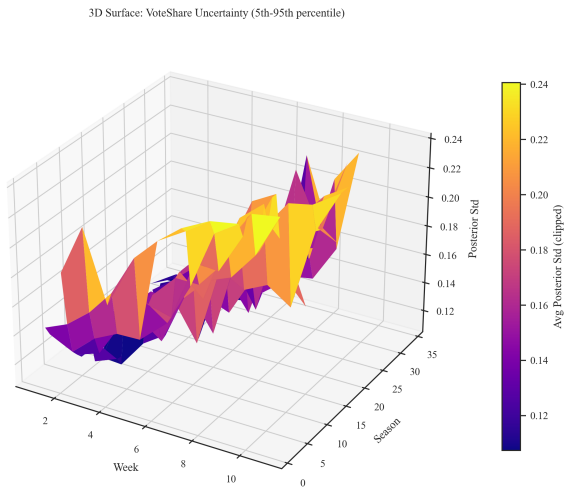
where gap$(\cdot)$ measures how far the sample is from producing the true elimination (e.g., for `Percent`, the eliminated candidates combined score above the minimum; for `Rank`-based rules, the eliminated candidates rank-sum below the maximum). We then resample $N_{\min}$ posterior samples according to $\{w_m\}$. This guarantees a stable posterior output for every usable season-week.

Posterior summaries. Given accepted samples $\{\mathbf{v}_{s,w}^{(k)}\}_{k=1}^{K}$, we report posterior mean and uncertainty (standard deviation) per couple:

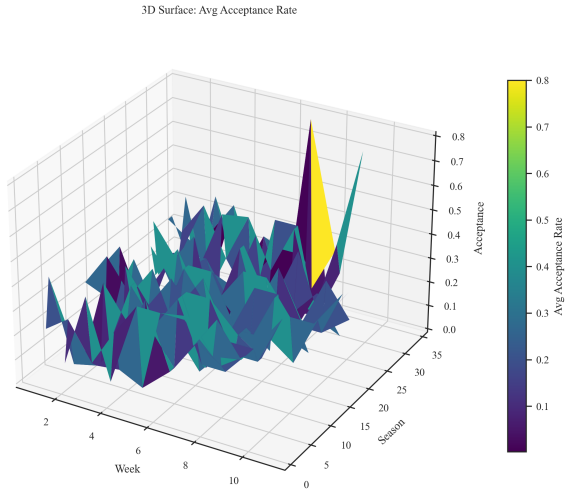
$$\mu_{s,w,i} = \frac{1}{K}\sum_{k=1}^{K} v_{s,w,i}^{(k)}, \qquad \sigma_{s,w,i} = \sqrt{\frac{1}{K}\sum_{k=1}^{K}\left(v_{s,w,i}^{(k)} - \mu_{s,w,i}\right)^2}.$$

4.1.4 Model Solution

We run the above inference for every season-week where: (i) a unique eliminated couple exists, (ii) at least two active couples have valid judge totals, and (iii) the eliminated couple is among the active set.



(a) 3D surface: vote-share uncertainty (5th–95th percentile, clipped).



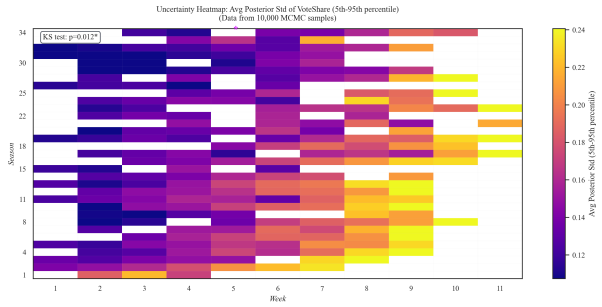
(b) 3D surface: average Phase-0 acceptance rate.

Figure 2: Task 1 diagnostic surfaces over the season–week grid.

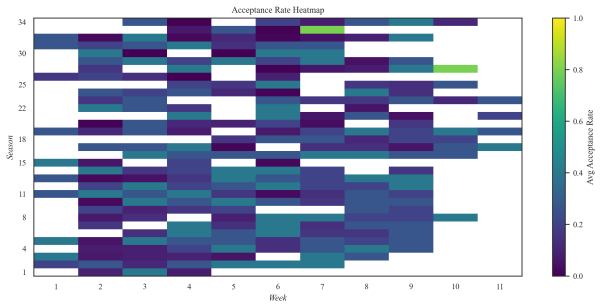
Diagnostics: uncertainty and inference difficulty. Because the observed data reveal only the *eliminated couple* (a coarse, discrete outcome), the latent fan shares $\mathbf{v}_{s,w}$ are generally *not identifiable*: many distinct vote-share vectors can produce the same elimination under the deterministic

rule. Consequently, a perfect replay of eliminations does *not* imply that the inferred fan shares are uniquely determined or "exact"; it only implies *consistency* with the rule.

We therefore report two complementary diagnostics that quantify how informative each week is for inference: (i) the posterior vote-share uncertainty (via `AvgVoteUncertainty`), and (ii) the Phase-0 rejection-ABC acceptance rate $\rho_{s,w}$, which serves as a proxy for identifiability. Intuitively, low acceptance means the set $\{\mathbf{v} : g_{\mathcal{R}}(\mathbf{J}, \mathbf{v}) = e\}$ occupies a small region of the simplex (a "thin" feasible set), so the elimination provides weak and/or highly boundary-dependent information. We visualize both quantities as season–week heatmaps and as 3D surfaces (clipped to the 5th–95th percentile range for readability). Figures 3a and 2a report vote-share uncertainty, while Figures 3b and 2b report Phase-0 acceptance rates.



(a) Week-level uncertainty heatmap (`AvgVoteUncertainty`).



(b) Acceptance-rate heatmap (higher ⇒ easier exact ABC matching).

Figure 3: Task 1 diagnostic heatmaps over the season–week grid.

Posterior predictive consistency (rule consistency, not unique identification). We verify that inferred fan shares reproduce observed eliminations by plugging the posterior mean $\boldsymbol{\mu}_{s,w}$ into the historical rule $\mathcal{R}_{\text{like}}(s)$:

$$\widehat{e}_{s,w}(\mathcal{R}_{\text{like}}) = g_{\mathcal{R}_{\text{like}}(s)}(\mathbf{J}_{s,w}, \boldsymbol{\mu}_{s,w}) .$$

Across 229 elimination weeks, the match rate is 1.00, with era breakdown: `Rank` (10/10), `Percent` (208/208), and `Bottom2+JudgeSave` (11/11). We emphasize that this replay check establishes internal consistency of the inference–rule pipeline; uncertainty and acceptance diagnostics above are needed to assess how tightly fan shares are constrained in each week.

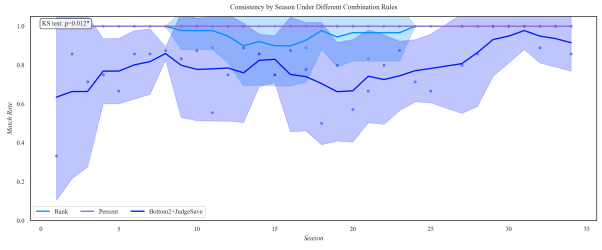


Figure 4: Consistency by season under different combination rules (using inferred fan shares).

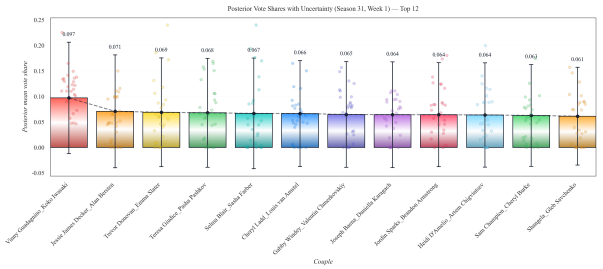


Figure 5: Example posterior fan vote shares with uncertainty bars (Top 12 shown).

Uncertainty of inferred fan votes. Per couple-week, posterior SD $\sigma_{s,w,i}$ has median 0.016 (IQR [0.000, 0.036]), 90th percentile 0.056, and maximum 0.346. We summarize week-level uncertainty by

$$\overline{\sigma}_{s,w} = \frac{1}{n_{s,w}} \sum_{i=1}^{n_{s,w}} \sigma_{s,w,i}$$

(exported as `AvgVoteUncertainty`). In our run, $\overline{\sigma}_{s,w}$ ranges from 0.000 to 0.253 (median 0.153, IQR [0.129, 0.189]), with the largest value at Season 13, Week 9.

Inference stability. Among 231 usable season-weeks, 228 achieved exact ABC acceptance ($\ell_{s,w} = 0$) and 3 required the bottom-2 relaxation ($\ell_{s,w} = 1$); none required Phase 2 soft resampling. The acceptance rate $\rho_{s,w}$ has median 0.200 (IQR [0.133, 0.333]), with range [0.001, 1.000].

Example posterior and downstream signals. Figure 4 summarizes season-level consistency under alternative rules (using inferred fan shares), and Figure 5 shows an illustrative posterior vote-share bar chart with uncertainty.

For Task 2, we export two summary views based on the same inferred fan shares: the flip rate between `Rank` and `Percent` (Figure 6) and the distribution of contestant-level survival-week shifts under `Percent` relative to `Rank` (Figure 7).

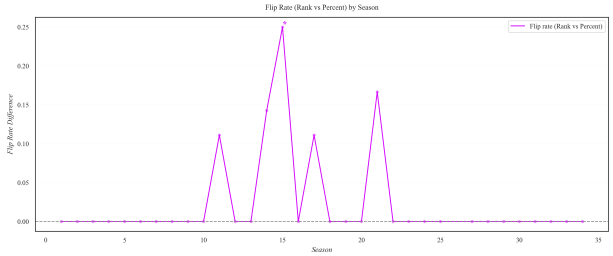


Figure 6: Flip rate between `Rank` and `Percent` by season (using inferred fan shares).

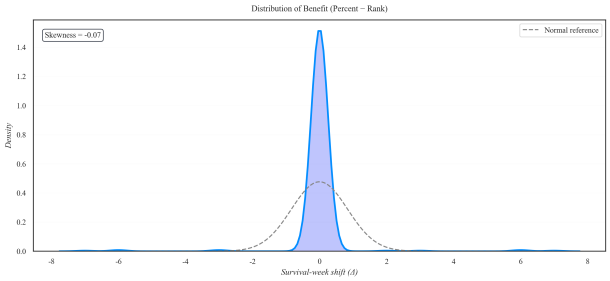


Figure 7: Distribution of contestant-level benefit (Percent − Rank) measured by survival-week shift Δ.

4.2 Modeling and Solving of Task 2: Voting-Rule Comparison and Who Benefits

4.2.1 Problem Analysis

Task 2 asks how outcomes change under different methods of combining judge scores and fan votes, and which contestants benefit or suffer under alternative rules. Using Task 1, we obtain posterior mean fan vote shares $\boldsymbol{\mu}_{s,w} = (\mu_{s,w,1}, \ldots, \mu_{s,w,n_{s,w}})$ for each season-week (s, w). We can therefore perform *counterfactual simulations* by applying different voting rules to the same judge-score vector $\mathbf{J}_{s,w}$ while holding fan shares fixed at $\boldsymbol{\mu}_{s,w}$.

In particular, we focus on: (i) rule-level consistency with observed eliminations, (ii) disagreement between rules (flip events), (iii) contestant-level survival-week shifts (Δ) measuring who benefits or suffers, and (iv) controversy cases where fan support and judge rankings appear to conflict.

4.2.2 Model Preparation

We consider three canonical voting rules implemented consistently across all seasons:

$$\mathcal{R} = \{\texttt{Rank}, \texttt{Percent}, \texttt{Bottom2+JudgeSave}\}.$$

Here, `Rank` aggregates ordinal ranks from judges and fans, `Percent` aggregates normalized score magnitudes, and `Bottom2+JudgeSave` introduces an additional judge intervention step that can override fan preference within the bottom-two set. Using Task 1, we plug in posterior mean fan shares $\boldsymbol{\mu}_{s,w}$ and evaluate

$$\widehat{e}_{s,w}(\mathcal{R}) = g_{\mathcal{R}}(\mathbf{J}_{s,w}, \boldsymbol{\mu}_{s,w})$$

as the counterfactual elimination outcome under rule $\mathcal{R}$.

4.2.3 Model Solution

Using the posterior mean fan vote shares from Task 1, we compute counterfactual eliminations under three rules: `Rank`, `Percent`, and `Bottom2+JudgeSave` (`Bottom2`). From these week-level outcomes $\widehat{e}_{s,w}(\mathcal{R})$, we summarize (i) match rates to historical eliminations, (ii) rule disagreement (flip events), and (iii) contestant-level survival-week shifts.

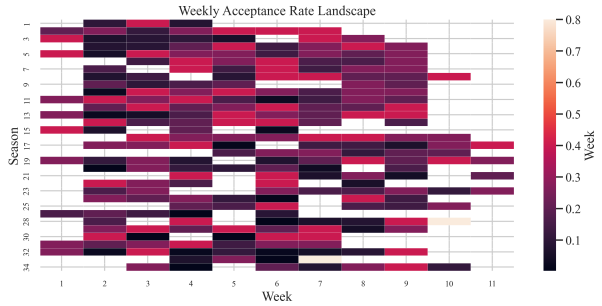


Figure 8: Task 2 diagnostic: ABC Phase-0 acceptance-rate heatmap (higher ⇒ easier exact matching).

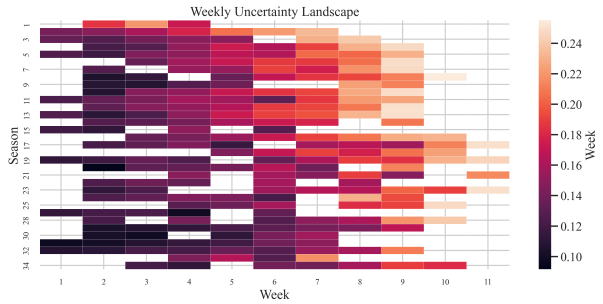
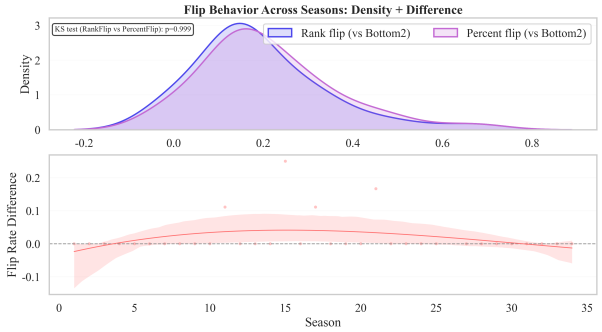


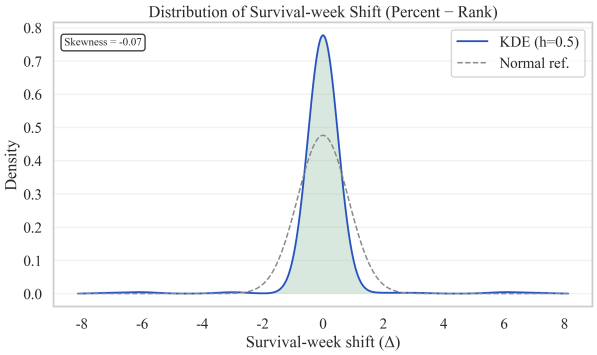
Figure 9: Task 2 diagnostic: uncertainty heatmap for inferred fan shares (week-level average posterior SD).

The pipeline exports:

- `method_outcomes_by_week.xlsx`: $\widehat{e}_{s,w}(\mathcal{R})$ and $\text{Match}_{\mathcal{R}}^{(s,w)}$ for each rule;
- `method_comparison_summary.xlsx`: season-level and overall consistency and flip summaries;
- `contestant_benefit_analysis.xlsx`: predicted elimination week $E_s(\mathcal{R}, \text{CoupleID})$ and survival-week shift ΔE for each pair of rules;
- `task2_week_consistency_diff.xlsx` and `task2_candidate_consistency_diff.xlsx`: detailed diagnostics including controversy cases and automatically detected sensitive cases.



(a) Flip behavior (Rank vs Percent).



(b) ECDF of $\Delta = E_s(\texttt{Percent}) - E_s(\texttt{Rank})$.

Figure 10: Task 2 rule-comparison summary (left to right): consistency by season under `Rank`/`Percent`/`Bottom2`; flip behavior between `Rank` and `Percent`; and the distribution of contestant-level survival-week shifts Δ.

We also report two week-level diagnostics from the post-processing run: ABC Phase-0 acceptance rates (Figure 8) and vote-share uncertainty (Figure 9).

We summarize the main Task 2 results in Figure 10: season-level match rates (panel a), Rank–Percent flip behavior (panel b), and the distribution of survival-week shifts (panel c). Overall, `Percent` is most consistent with historical eliminations, `Rank` is slightly lower, and `Bottom2` can be lower due to the judge-save override. In producer terms: `Percent` is the clearest "fair aggregation" default, while `Bottom2+JudgeSave` is better viewed as a deliberate *show-format* choice that trades fairness for drama and judge authority.

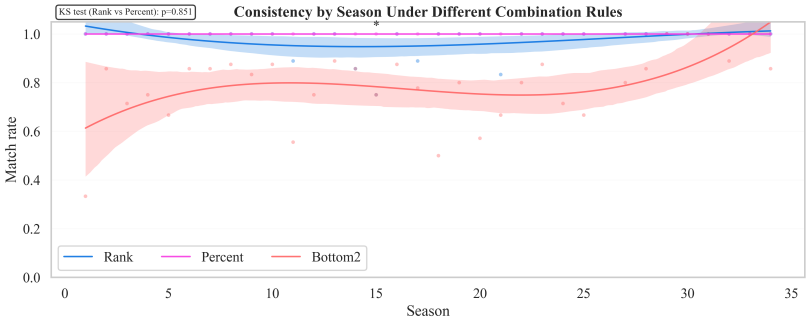


Figure 11: Consistency by season.

We define a flip event by

$$\text{Flip}^{(s,w)}_{\texttt{Rank,Percent}} = \mathbb{I}\left[\widehat{e}_{s,w}(\texttt{Rank}) \neq \widehat{e}_{s,w}(\texttt{Percent})\right],$$

and measure contestant-level impact via

$$\Delta = E_s(\texttt{Percent}) - E_s(\texttt{Rank}).$$

So $\Delta > 0$ means the contestant survives longer under `Percent`. Figure 10c (also Figure 10b) shows the ECDF of Δ; a distribution concentrated around 0 indicates that, in aggregate, neither

rule systematically advantages contestants, although individual contestants can still experience large positive or negative shifts.

The prompt lists four high-profile controversy cases, which we evaluate to see whether changing the combination method (`Rank` vs `Percent`) would change the outcome, and how adding `Bottom2+JudgeSave` would impact results. Table 2 reports each contestant's predicted elimination week under each rule.

Season	Celebrity	Rank	Percent	Bottom2	Bottom2-Percent
2	Jerry Rice	12	12	7	-5
4	Billy Ray Cyrus	8	8	7	-1
11	Bristol Palin	12	12	8	-4
27	Bobby Bones	12	12	12	0

Table 2: Given controversy cases: predicted elimination week under different rules.

From Table 2 we conclude that switching between `Rank` and `Percent` does not change any of the four listed outcomes, while introducing `Bottom2+JudgeSave` eliminates three of the four contestants earlier (Jerry Rice: -5 weeks; Bristol Palin: -4 weeks; Billy Ray Cyrus: -1 week). Bobby Bones (Season 27) remains unchanged under all three rules, suggesting his fan advantage is strong enough that even a judge-save mechanism does not reverse the season-level trajectory.

Beyond the hand-picked controversy cases, we identify contestants whose predicted elimination week changes most across rules. Table 3 lists the top five automatically detected method-sensitive cases.

Season	Celebrity	ProDancer	WeirdScore	Rank	Percent	Bottom2
3	Shanna Moakler	Jesse DeSoto	12.5	2	2	12
13	Elisabetta Canalis	Valentin Chmerkovskiy	12.5	2	2	12
1	Kelly Monaco	Alec Mazo	12.5	12	12	2
20	Redfoo	Emma Slater	12.5	2	2	12
1	Trista Sutter	Louis van Amstel	12.5	2	2	12

Table 3: Top 5 most method-sensitive contestants (automatically detected).

These cases highlight that `Bottom2+JudgeSave` can produce dramatic outcome shifts even when `Rank` and `Percent` agree. Therefore, introducing the judge-save rule changes the decision structure qualitatively: it is not a minor tweak, but a mechanism that can override fan votes in the bottom-two weeks.

To make disagreements concrete, Table 4 shows example weeks where `Rank` and `Percent` predict different eliminated couples.

Recommendation (producer-facing). If the primary goal is *fair and transparent aggregation* of judges and fans, we recommend **`Percent`**: it combines normalized magnitudes, preserves intensity of support, and is straightforward to communicate on-air. If the primary goal is *live-show drama and judge authority*, then **`Bottom2+JudgeSave`** is a viable format choice, but it should be

Season	Week	Actual Eliminated	Pred_Rank	Pred_Percent
11	6	Audrina Patridge_Tony Dovolani	Kyle Massey_Lacey Schwimmer	Audrina Patridge_Tony Dovolani
14	9	Maria Menounos_Derek Hough	William Levy_Cheryl Burke	Maria Menounos_Derek Hough
15	6	Sabrina Bryan_Louis van Amstel	Kelly Monaco_Valentin Chmerkovskiy	Sabrina Bryan_Louis van Amstel
17	5	Christina Milian_Mark Ballas	Amber Riley_Derek Hough	Christina Milian_Mark Ballas
21	9	Alexa PenaVega_Mark Ballas	Carlos PenaVega_Witney Carson	Alexa PenaVega_Mark Ballas

Table 4: Example weeks where `Rank` and `Percent` predict different eliminations (first rows shown).

adopted explicitly as an entertainment mechanism: it can materially change outcomes by overriding fan votes in bottom-two weeks and may amplify perceived bias toward judges (Tables 2–3).

4.3 Modeling and Solving of Task 3: Explanatory Regression and Sensitivity

4.3.1 Problem Analysis

Task 3 asks us to explain why certain contestants survive longer, receive higher judge scores, or obtain stronger fan support, and to assess whether these conclusions are robust to modeling choices. Using the dataset, we define three outcome targets that correspond directly to the three perspectives of the show: (i) **performance** measured by `WeeksSurvived`, (ii) **judges** measured by `AvgJudgeTotal`, and (iii) **fans** measured by `AvgFanShare`. We fit an explanatory regression model that quantifies the marginal effects of interpretable attributes while controlling model complexity through regularization.

4.3.2 Model Preparation

We construct a contestant-level modeling table by aggregating weekly information within each season. Each row represents one contestant-couple (identified by `CoupleID`) and includes:

- Numeric: `Age`;
- Categorical: `Partner` (pro dancer), `Industry`, `HomeState`, `HomeCountry`;
- Targets: `WeeksSurvived`, `AvgJudgeTotal`, `AvgFanShare`.

To ensure comparability across targets, we fit separate models for each target using the same feature set, dropping rows with missing values for the specific target being modeled.

To avoid extremely high-dimensional one-hot encodings, each categorical variable is capped to its Top-K most frequent categories and all remaining categories are grouped into an `Other` class. This makes the regression stable and interpretable.

4.3.3 Model Construction

Let $\boldsymbol{x}_i \in \mathbb{R}^d$ denote the one-hot encoded feature vector (plus standardized age) for contestant i, and let y_i denote one of the three targets. We use a linear model with intercept:

$$\hat{y}_i = \beta_0 + \boldsymbol{x}_i^\top \boldsymbol{\beta}.$$

To control overfitting and stabilize coefficients under correlated categorical predictors, we apply ridge regression:

$$(\hat{\beta}_0, \hat{\boldsymbol{\beta}}) = \arg\min_{\beta_0, \boldsymbol{\beta}} \sum_{i=1}^{n} (y_i - \beta_0 - \boldsymbol{x}_i^\top \boldsymbol{\beta})^2 + \lambda \|\boldsymbol{\beta}\|_2^2.$$

Model selection is performed using K-fold cross-validation (default $K = 5$). We sweep over λ (ridge strength) and the Top-K caps, selecting the configuration that maximizes predictive stability (tie-breaking by better cross-validated performance).

To summarize feature importance in an interpretable way, we compute grouped effect magnitudes (absolute standardized coefficient summaries) for:

$$\{\texttt{Age}, \texttt{Partner}, \texttt{Industry}, \texttt{HomeState}, \texttt{HomeCountry}\},$$

and compare their impacts across the three targets.

4.3.4 Model Solution

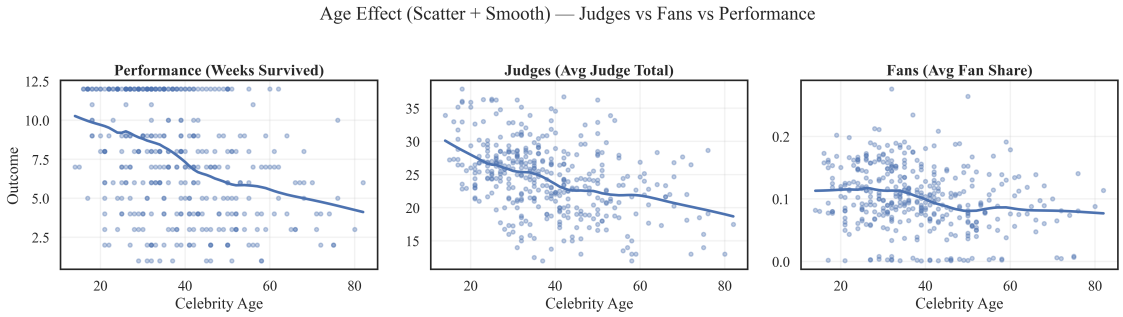


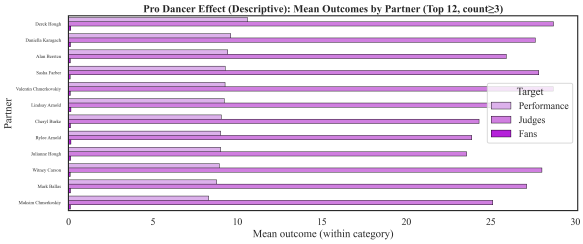
Figure 12: Supplementary diagnostic: age vs. outcomes with a smooth trend line (Task 3).

The Task 3 pipeline generates both predictive and explanatory outputs. All modeling results are exported in `task3_hyperparam_sweep_and_effects.xlsx`, which contains the cleaned contestant-level modeling table and cross-validation sweep results over (λ, K) settings.

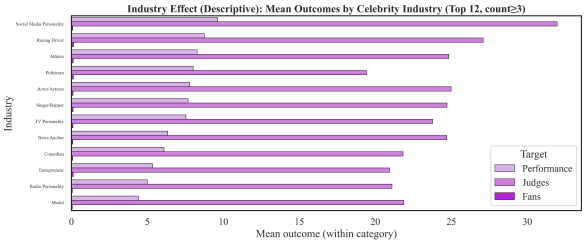
To summarize the findings in the main text, we report three core visual diagnostics: (i) a grouped impact heatmap across features and targets, (ii) a grouped-bar comparison of feature importance across targets, and (iii) the cross-validated R^2 sensitivity curve as the ridge penalty varies.

We additionally visualize the marginal relationship between `Age` and the modeled outcomes using a scatter plot with a smooth trend line (Figure 12).

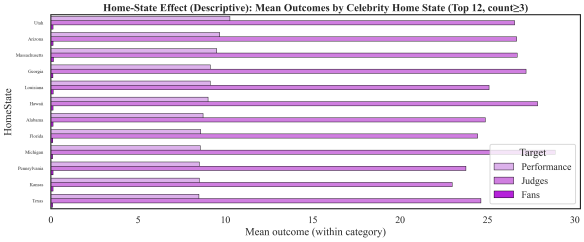
Supplementary descriptive breakdowns. To complement the ridge coefficients with an intuitive sanity check, we also report descriptive subgroup means (Top 12 categories, count$\geq$ 3) for the



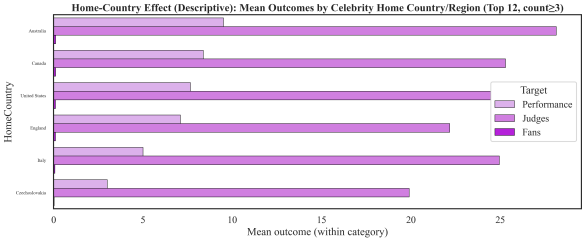
(a) Partner (Top 12, count≥ 3).



(b) Industry (Top 12, count≥ 3).



(c) HomeState (Top 12, count≥ 3).



(d) HomeCountry/Region (Top 12, count≥ 3).

Figure 13: Task 3 supplementary descriptive plots: mean outcomes by subgroup (Top 12 categories, count≥ 3). These provide an intuitive cross-check that aligns with the regression-based importance ordering.

main high-cardinality attributes. Figure 13 shows that partner-level differences are the most pronounced (especially for judge outcomes), while industry and geographic attributes exhibit smaller but visible separation, consistent with the grouped-impact results.

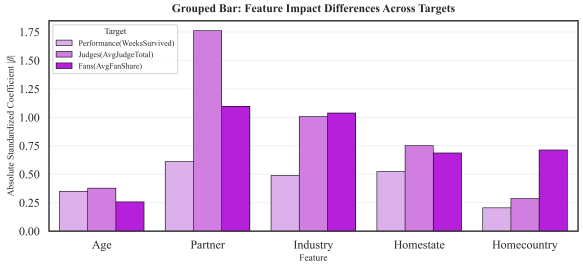


Figure 14: Feature-importance comparison across targets (grouped coefficient magnitudes).

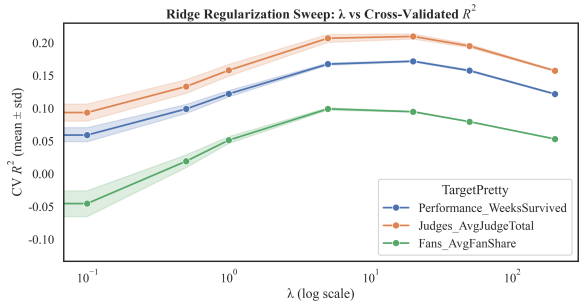


Figure 15: Sensitivity of CV R^2 to λ.

Figures 17–14 highlight a clear separation between judge-driven and fan-driven outcomes. The grouped-impact heatmap (Figure 17) indicates that `Partner` is the strongest explanatory factor for `AvgJudgeTotal`, and it remains influential for `AvgFanShare`, consistent with the idea that professional dancer pairing affects both scoring and audience support. Meanwhile, `Industry` and regional attributes (`HomeState`, `HomeCountry`) show moderate but non-negligible signals, suggesting that fan engagement depends partly on contestant identity and background beyond judge performance.

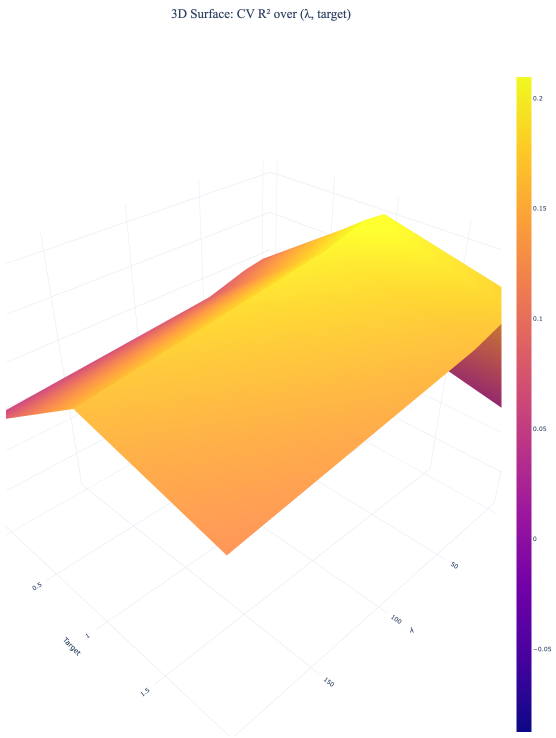


Figure 16: CV surface over λ across targets.

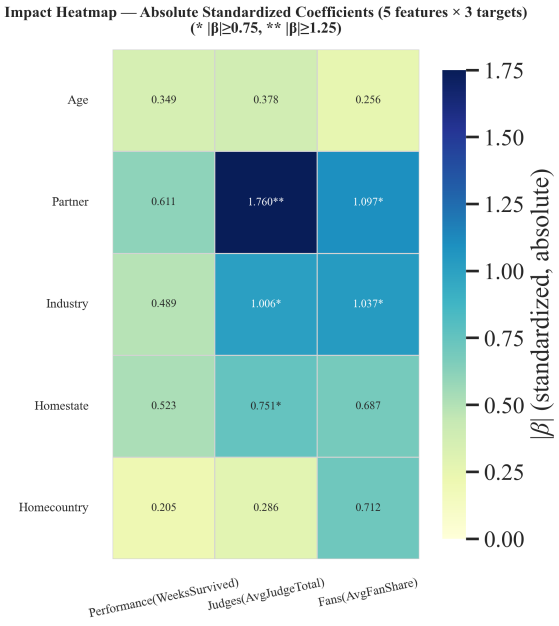


Figure 17: Grouped impact heatmap (|standardized coefficients|) across features and targets.

The ridge sweep in Figure 15 shows that predictive performance is relatively stable across a broad range of λ, which supports the robustness of the qualitative conclusions to regularization strength. For additional descriptive breakdowns by category level (Partner/Industry/HomeState/HomeCountry), we provide supplementary figures in the Appendix to avoid overloading the main narrative.

4.4 Modeling and Solving of Task 4: FairVote Backtest, Tuning, and Robustness

4.4.1 Problem Analysis

Task 4 requires proposing and validating an improved voting mechanism that balances: (i) agreement with historical eliminations (consistency), (ii) entertainment value (measured by "close-call" triggers / gate rate), and (iii) fan protection (how fan-preferred contestants are treated), while also monitoring potential feature-driven bias (via a Bias Stability Index, BSI).

4.4.2 Model Preparation

Inputs from Tasks 1–2. We use week-level judge scores and inferred fan shares from Task 1 as the baseline signals. Task 2 provides a reference set of outcomes under existing rules.

FairVote tunable parameters. The backtest report indicates a two-parameter tuning surface:

- k: an uncertainty penalty (controls how strongly vote uncertainty affects the rule),
- τ: a "close-call" gate threshold (controls when special handling is triggered).

The tuning sweep records metrics such as `Consistency`, `ExcitementRate`, and `DeltaFanShareElim(FV-Actu` optionally also `BSI_FV_sumAbsGamma` and `R2_FV`.

4.4.3 Model Construction

Backtest objective metrics. For each parameter pair (k, τ), the model generates week-level predictions $\widehat{E}_{\mathrm{FV}}^{(s,w)}$ and computes:

$$\begin{aligned}
\text{Consistency}(k,\tau) &= \frac{1}{W}\sum_{(s,w)} \mathbb{I}\Big[\widehat{E}_{\mathrm{FV}}^{(s,w)} = e_{s,w}\Big], \\
\text{ExcitementRate}(k,\tau) &= \frac{1}{W}\sum_{(s,w)} \mathbb{I}[\text{Gate triggered in week } (s,w)], \\
\Delta\text{FanShareElim}(k,\tau) &= \frac{1}{W}\sum_{(s,w)} \Big(\mu_{s,w,\widehat{E}_{\mathrm{FV}}^{(s,w)}} - \mu_{s,w,e_{s,w}}\Big),
\end{aligned}$$

where negative ΔFanShareElim indicates that FairVote tends to eliminate contestants with *lower* fan share than the historical elimination (more fan-protective).

Parameter selection. We select $(k^\star, \tau^\star)$ as the best trade-off point from the sweep report, using maximum Consistency as the primary objective and reporting the associated ExcitementRate and fan-protection metric. (Implementation uses a best-parameter sheet `best_params` in the report.)

Robustness checks. The sensitivity script further performs [4]:

- **Surface visualization**: heatmaps over (k, τ) for Consistency / ExcitementRate / fan-protection, with the selected best point marked;
- **Bootstrap confidence intervals**: resampling across weeks to provide CIs for key metrics at $(k^\star, \tau^\star)$;
- **LOSO stability (leave-one-season-out)**: verifying that the results are not driven by any single anomalous season.

4.4.4 Model Solution

We backtest `FairVote` over a (k, τ) grid and summarize performance and robustness.

Fairness definition (operational). We define "fairness" as a multi-objective criterion evaluated from three complementary angles: (i) *fan protection*, measured by the eliminated-couple fan-share shift $\Delta = \mu_{\text{elim}}^{\text{FV}} - \mu_{\text{elim}}^{\text{Actual}}$ (more protective $\Leftrightarrow$ smaller/negative Δ); (ii) *volatility control*, measured by the excitement (gate-trigger) rate to avoid excessive rule intervention; and (iii) *bias control*, measured by the Bias Stability Index (BSI), where lower indicates weaker feature-driven

bias. Thus, `FairVote` is a tunable policy knob: (k, τ) controls the trade-off among historical consistency, intervention frequency, fan protection, and bias stability.

Full sweep outputs, the selected best parameters, week-level predictions, and season summaries are exported to `fairvote_backtest_report.xlsx`. Bootstrap CIs and leave-one-season-out (LOSO) stability checks are reported in `task4_sensitivity_report.xlsx` (figures in `results/figures_task4_sensitivity/`).

Metric	**Value**
Consistency (match rate)	**0.355** [0.294, 0.416]
Excitement rate (gate)	0.255
Fan-favoring Δ mean (all)	0.114
Fan-favoring Δ median (all)	0.105
$\mathbb{P}(\Delta > 0)$ (all)	0.645
Fan-favoring Δ mean (gate-only)	0.125
Gate weeks (count)	59
ω (mean / median)	0.350 / 0.350
BSI (FairVote) $\sum \lvert\gamma_k\rvert$	3.842
BSI baseline (Rank/Percent/Bottom2)	1.513 / 1.500 / 1.720
Best params (k, τ)	2.000, 0.200
ω bounds $(\omega_{\min}, \omega_{\max})$	0.350, 0.650

Table 5: FairVote backtest summary at the selected operating point. The match-rate consistency is reported with a 95% bootstrap CI.

Table 5 reports the selected operating point $(k^\star, \tau^\star) = (2.0, 0.2)$. At this setting, the match-rate consistency is 0.355 with a 95% bootstrap CI [0.294, 0.416], and the gate (c̈lose-call¨) rate is 0.255 (59 gate weeks), indicating moderate use of the special-handling mechanism.

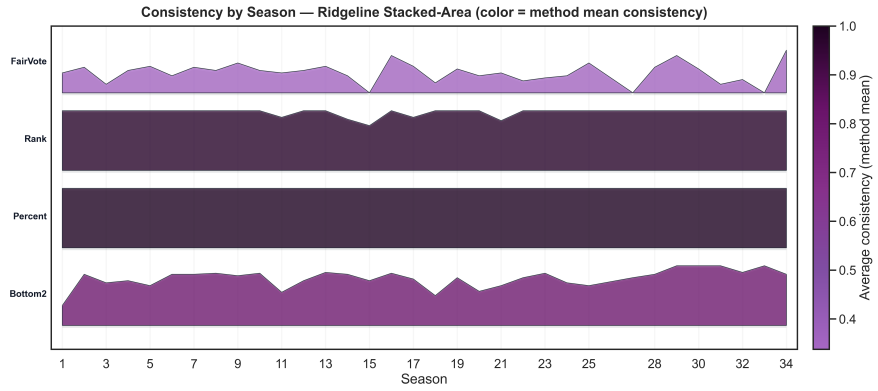
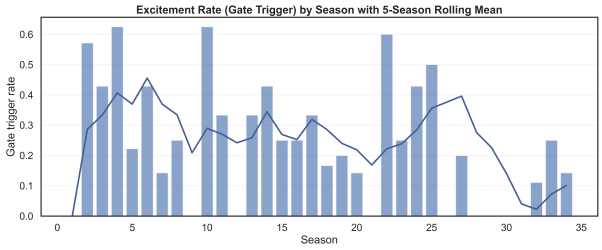


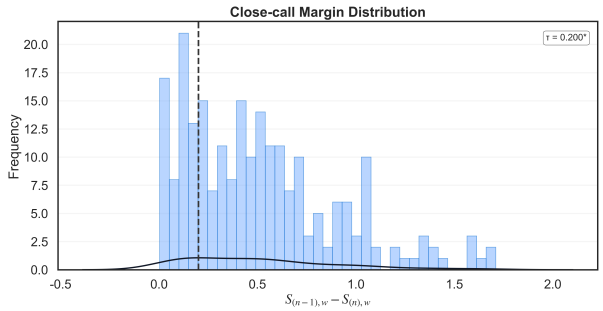
Figure 18: Season-level consistency (match rate) comparison for `FairVote` versus baseline rules. Color indicates each method's average consistency.

To interpret the gate mechanism, Figure 19a shows the gate-trigger rate by season (with a

rolling mean), and Figure 19b shows the close-call margin distribution $S_{(n-1),w} - S_{(n),w}$ with the threshold $\tau^\star = 0.2$ marked.



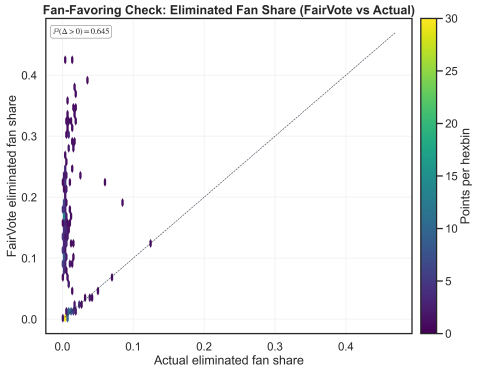
(a) Gate-trigger (excitement) rate by season with rolling mean.



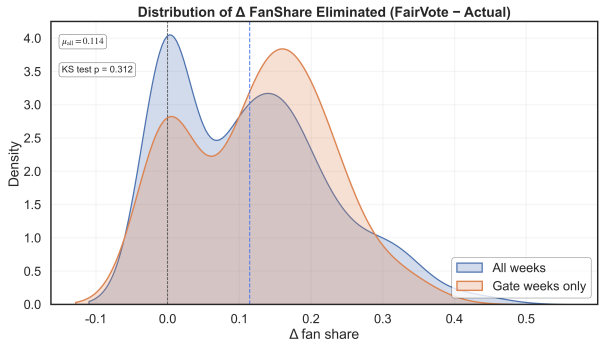
(b) Close-call margin distribution; dashed line indicates $\tau^\star = 0.2$.

Figure 19: Task 4 diagnostics for the close-call mechanism: how often the gate is triggered across seasons and how the trigger threshold $\tau^\star$ relates to the empirical margin distribution.

Figure 18 compares season-level consistency for `FairVote` against the three baseline rules. `Rank` and `Percent` remain the most consistent overall; `Bottom2` is lower due to judge overrides. `FairVote` achieves moderate consistency with noticeable season-to-season variation, reflecting its trade-off between matching history and the gate/fan-protection objectives.



(a) Week-level eliminated fan share: FairVote vs actual.



(b) Distribution of $\Delta = \mu^{\text{FV}}_{\text{elim}} - \mu^{\text{Actual}}_{\text{elim}}$ (all weeks vs gate-only).

Figure 20: FairVote fan-protection diagnostics at $(k^\star, \tau^\star)$.

Fan protection is assessed by comparing the eliminated couple's fan share under FairVote to the historical elimination. Figure 20 shows week-level eliminated fan shares (hexbin) and the distribution of $\Delta = \mu^{\text{FV}}_{\text{elim}} - \mu^{\text{Actual}}_{\text{elim}}$ (density). At $(k^\star, \tau^\star)$, $\mathbb{P}(\Delta > 0) = 0.645$, so FairVote more often eliminates a higher-fan-share couple than the historical outcome; we therefore treat fan protection as a tunable trade-off controlled by (k, τ).

Robustness results (metric surfaces, bootstrap CIs, and LOSO stability) are provided in `task4_sensitivity` indicating the chosen operating point is not driven by a single season.

5 Sensitivity Analysis

5.1 Sensitivity Analysis of Task 1: Bayesian Fan-Vote Inference

We test robustness of the Stable-ABC fan-vote inference to two key hyperparameters: the Dirichlet prior concentration α and the rejection-ABC budget `MAX_DRAWS`. We track mean acceptance rate and mean fallback rate.

Figure 21 shows that acceptance changes only mildly over $\alpha \in [0.1, 1.0]$, while fallback quickly approaches 0 once $\alpha \geq 0.3$, indicating stable inference under moderately informative priors. Figure 22 suggests diminishing returns from increasing `MAX_DRAWS`, since fallback remains near 0 and acceptance improves only marginally. Overall, Stable-ABC exhibits stable behavior under reasonable hyperparameter perturbations.

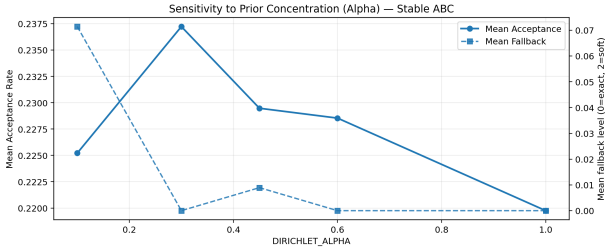


Figure 21: Sensitivity to Dirichlet prior concentration α: mean acceptance rate and fallback rate.

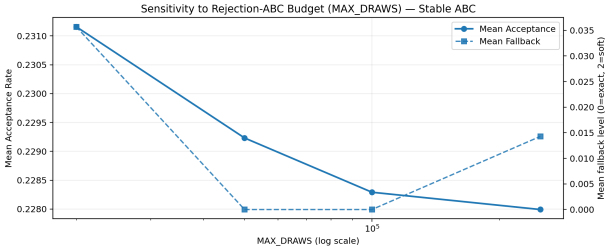
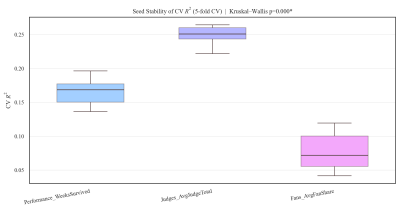


Figure 22: Sensitivity to rejection-ABC budget `MAX_DRAWS`: mean acceptance rate and fallback rate.

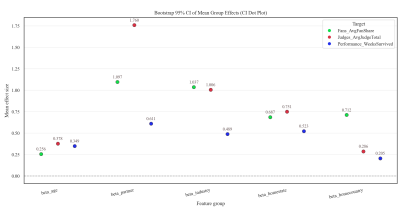
5.2 Sensitivity Analysis of Task 3: Explanatory Regression and Sensitivity

We assess robustness of the Task 3 ridge-regression explanations from three aspects: (i) seed stability under different cross-validation splits, (ii) bootstrap stability of grouped effect magnitudes, and (iii) sensitivity to ridge regularization strength λ.

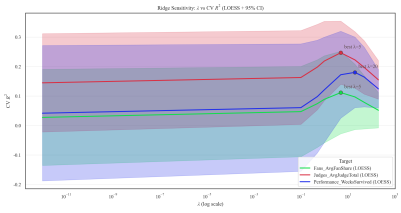
As shown in Figure 23a, the 5-fold CV R^2 distributions are stable across seeds, with `Judges_AvgJudgeTotal` consistently achieving the highest explainability, followed by `Performance_WeeksSurvived`, while `Fans_AvgFanShare` remains the most difficult target. Figure 23b further confirms that the grouped-effect ranking is robust under bootstrapping: `Partner` is the dominant factor for judge outcomes and remains influential for fan share, while `Industry` and geographic features contribute weaker but nonzero signals.



(a) Seed stability of 5-fold CV R^2 across targets.



(b) Bootstrap 95% CI of mean grouped effects.



(c) CV R^2 sensitivity to ridge λ (log scale).

Finally, Figure 23c shows smooth and bounded changes of CV R^2 over a wide range of λ, indicating that conclusions are not driven by a fragile regularization choice. Overall, Task 3 results are robust to random-seed perturbations, resampling uncertainty, and ridge-penalty tuning.

5.3 Sensitivity Analysis of Task 4: FairVote Backtest, Tuning, and Robustness

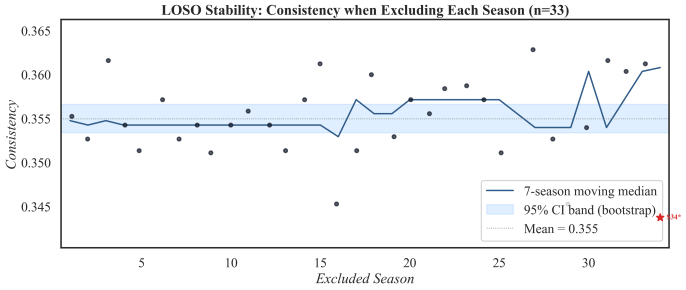
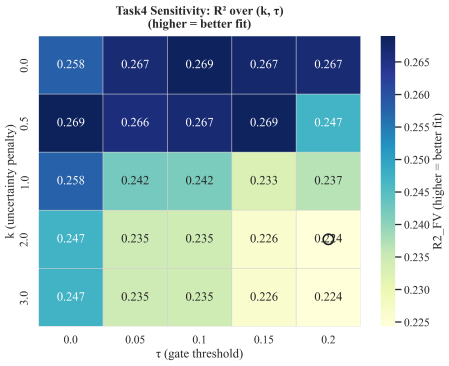
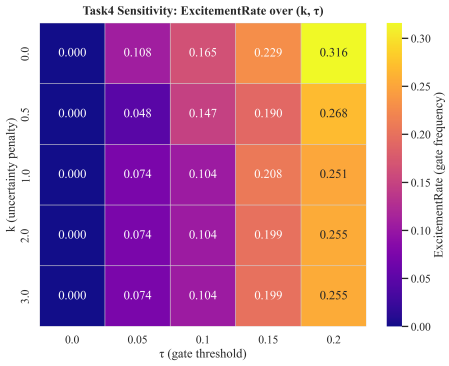


Figure 24: LOSO stability of consistency across seasons.

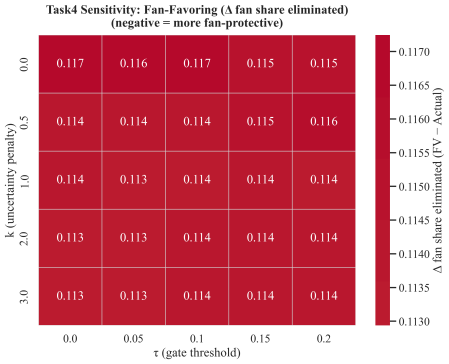
We evaluate robustness of `FairVote` over (k, τ) using five diagnostics. Figure 25 summarizes four sensitivity surfaces (R^2, excitement rate, fan-impact, and BSI), and Figure 24 reports leave-one-season-out (LOSO) stability. The surfaces are smooth with no sharp optima, suggesting moderate perturbations of (k, τ) do not qualitatively change conclusions.



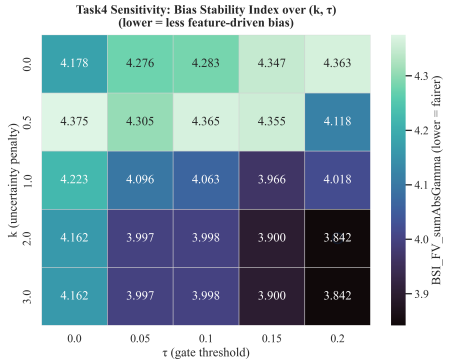
(a) R^2 over (k, τ).



(b) Excitement rate (gate frequency).



(c) Fan-impact: Δ fan share eliminated (FV − Actual).



(d) BSI (lower ⇒ less feature-driven bias).

Figure 25: Task 4 sensitivity surfaces over (k, τ): R^2, excitement rate, fan-impact, and BSI.

6 Model Evaluation

6.1 Advantages

- **Reproducible pipeline.** A modular workflow (Task 1 inference → Task 2 replay → Task 3 explanation → Task 4 design) with all key outputs exported for auditing.
- **Internal validity via replay.** Using posterior means $\boldsymbol{\mu}_{s,w}$, the historical-rule replay matches eliminations in 229 weeks (match rate 1.00), verifying correct rule implementation.
- **Uncertainty + identifiability diagnostics.** Alongside $\mu_{s,w,i}$ we report $\sigma_{s,w,i}$ and week-level signals (`AvgVoteUncertainty`, acceptance rate $\rho_{s,w}$) to flag hard-to-identify weeks.
- **Stable inference.** The three-phase ABC rarely needs relaxation: 228/231 weeks succeed with exact matching, 3 use bottom-2 relaxation, none require soft resampling.
- **Actionable counterfactuals and interpretable explanations.** Holding fan votes fixed enables clean rule comparisons (flip events, survival shifts Δ). Ridge regressions provide interpretable, regularized feature effects.
- **Transparent rule design with robustness checks.** `FairVote` turns fairness into a tunable trade-off over (k, τ) and is supported by grid sweeps, bootstrap CIs, and LOSO stability.

6.2 Limitations

- **Non-identifiability of fan votes.** Eliminations are a coarse signal; many vote-share vectors can yield the same eliminated couple, so replay confirms rule-consistency, not uniqueness.
- **Prior and ABC design sensitivity.** The posterior depends on α and on relaxation choices (e.g., bottom-2 membership), especially in low-acceptance weeks.
- **Rule simplification risk.** The era-based rules approximate broadcast logic; unmodeled special weeks (e.g., double eliminations/withdrawals) may affect comparability.
- **Explanations are correlational.** Ridge coefficients capture associations; omitted factors (themes, injuries, narrative effects) can confound interpretation.
- **FairVote trade-offs are explicit, not universally "fair".** At $(2.0, 0.2)$, $\mathbb{P}(\Delta > 0) = 0.645$, so fan protection is not guaranteed; different objective weights would select different (k, τ).

7 Memo: Fair and Exciting Elimination Design: Rank vs Percent vs Judge Save

To: Producer, *Dancing with the Stars* (DWTS)
From: Team #2608125
Date: 02/02/2026
Subject: Fair and Exciting Elimination Design: Rank vs Percent vs Judge Save

Purpose. Judge scores are public while fan votes are hidden, so eliminations only partially reveal audience support. We built a data-driven framework to estimate weekly fan support with uncertainty, compare rule outcomes, explain key drivers, and test a tunable rule design.

Data. We use season-week judge totals and keep weeks with a single confirmed elimination. Historical formats are grouped into three rule types: `Rank`, `Percent`, and `Bottom2+JudgeSave`.

Finding 1: Fan support is inferable but not uniquely recoverable. Since only the eliminated couple is observed, many vote patterns can produce the same result. Our method estimates fan-support levels and reports uncertainty to flag hard-to-identify weeks.

Evidence (stability). Replaying eliminations using inferred fan support matches history in **229/229** elimination weeks. Across **231** usable weeks, almost all succeed under strict matching, showing the pipeline is stable for analysis.

Finding 2: Rule choice mainly matters in close weeks. Holding fan support fixed, rule replay shows disagreements between: `Rank` (order-based), `Percent` (magnitude-based), and `Bottom2+JudgeSave` (judge override, higher shock potential).

Recommendation. **`Percent`** is the best default for fairness and clarity. **`Bottom2+JudgeSave`** increases drama and judge control, but can override fan preference.

Finding 3: What drives outcomes. Regression results suggest the **professional partner** is the strongest driver of judge scores and remains influential for fan support.

Finding 4: FairVote provides a tunable compromise. `FairVote` can trade off fairness vs excitement using adjustable parameters. At our selected setting, it triggers "close-call" handling in about one quarter of elimination weeks.

Limitations. Fan support is inferred from eliminations only, and results depend on correct handling of special-week rule exceptions.

Bottom line. Use `Percent` for fairness and transparency; use Judge Save for producer-driven drama. `FairVote` offers a controllable middle option.

8 References

[1] M. A. Beaumont, W. Zhang, and D. J. Balding, "Approximate Bayesian computation in population genetics," *Genetics*, 162(4), 2002.

[2] A. Gelman, J. Carlin, H. Stern, et al., *Bayesian Data Analysis*, 3rd ed., CRC Press, 2013.

[3] S. A. Sisson, Y. Fan, and M. Beaumont, *Handbook of Approximate Bayesian Computation*, CRC Press, 2018.

[4] A. Saltelli, S. Tarantola, and K. P.-S. Chan, "Sensitivity analysis for importance assessment," *Risk Analysis*, 20(2), 2002.

[5] L. Prandtl, Fluid motions with very small friction, Proceedings of the 3rd International Mathematical Congress, Heidelberg: H. Schlichting, 1904, 484-491.

[6] A. E. Hoerl and R. W. Kennard, "Ridge regression: Biased estimation for nonorthogonal problems," *Technometrics*, 12(1), 1970, pp. 55–67.

[7] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed., Springer, 2009.

[8] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI)*, 1995.