

To Make a Difference, Comparative Gain, and Power Play Optimization



Hammer-Time Analytics CSAS 2026 – Curling

Abstract:

In Mixed Doubles Curling, the Power Play is a strategic option available to the team in possession of the Hammer (last stone of the game) only one time per game. Of the questions marked by the 2026 Connecticut Sports Analytics Symposium, choosing the optimal time to utilize the Power Play was the most appealing. Our primary approach utilizes a Random Forest classifier (AUC 0.89) to model the Win Probability, P_{win} , as a function of game state. By introducing a new metric, Win Probability

Added (WPA), we can quantify the nuanced strategic value of the Power Play. Our analysis reveals that the Power Play is used as a catch-up tactic best deployed between Ends 5-7 when trailing by 2-4 points (WPA increases by +26%). Interestingly, our model found that invoking the Power Play when

leading yields a negative WPA value. Because of an increase in variance and risk, we do not recommend using the Power Play when your team is leading. Lastly, our modeling investigated “Shot

Selection” strategy and found a null result ($\Delta x < 1$ in.) between both winning and losing stone guarding placements. This suggests that execution (shots), rather than defensive positioning, is the predominant driver of desired game outcomes (wins).

Table of Contents

1. Introduction

- 1.1 – Introduction to Curling (Power Play, Hammer, Stones, House)
- 1.2 – Global considerations for modeling

2. Modeling Methodology

- 2.1 – Assumptions & Justifications
- 2.2 – Model Selection
- 2.3 – Defining our Variables
- 2.4 – Data & Feature Engineering

3. Strategic Analysis of the Power Play (when should teams invoke it?)

- 3.1 – Twin Earths Simulation
- 3.2 – Testing
- 3.3 – Results

4. Execution Analysis of the Power Play (how should teams invoke it?)

- 4.1 – “Magic Spot” Hypothesis
- 4.2 – Testing
- 4.3 – Results

5. Limitations and Improvements

- 5.1 – Additional Considerations
- 5.2 – Comparative Gain Notes (Monte Carlo for opportunity cost)

6. Conclusion

7. References

8. Appendix

1. Introduction

1.1 Curling

Curling is a sport that is most widely known for its event at the Winter Olympics. Players slide 40-pound granite stones across a long sheet of ice into a circular target called the House. Each team consists of four players who alternate between delivering stones and sweeping. The curler launches from their House and delivers the stone across the sheet. Sweepers use brooms to help control a stone's speed and path.

An End is completed when 16 stones are delivered, and whichever team ends with their stone closest to the center of their opponents' House scores a point. The Hammer is the strategic advantage the last team to throw holds. The team that holds the Hammer gets to finalize the outcome of the End and can also invoke a Power Play.

The Power Play is a special rule only available in Mixed Doubles, where the team in possession of the Hammer can change the position of the pre-placed defensive stones at the start of an End to create a strategic advantage. Without the Power Play, two stones begin in the center of the sheet, one in the House and another as the center guard — but with, these stones are moved to one side so that the in-House stone sits on the tee line protected by a corner guard.

1.2 Global Modeling Considerations

To answer the question: “When is the most advantageous moment to use the Power Play?” We broke the big question into smaller parts. Firstly, what is the point of using a Power Play? To win the match. Secondly, what factors contribute to a win?

Our team found four primary factors that contribute to a Curling team's win.

1. **Points** – The team with the most points wins the match. We initially analyzed when teams earn their points and if a Power Play was used.
2. **Comparative gain** – Our team created this term to explain the following: relative to a team's total points in a moment of time, how much does a Power Play add to it? When does this occur? What state is the team using the Power Play in (winning or losing)?
3. **Rock(s) location** – Another advantage to a Power Play is that all the stones are moved into a certain position, which could either be beneficial (in most cases it is) or disadvantageous (if the team is in an already-comfortable position).

Pulling from the first definer, points, we first began predicting the effect on points by the Power Play with linear regression in R. By modelling the following: `lm(result ~ PowerPlayUsed)`, we noticed a significant advantage correlated with using the Power Play. A coefficient of +0.71 points per End ($p < 2e^{-16}$) was given.

```
##
## Call:
## lm(formula = Result ~ PowerPlayUsed, data = ends)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.6421 -0.9333 -0.6421  0.0667  8.0667
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.93328    0.02309   40.41  <2e-16 ***
## PowerPlayUsed  0.70886    0.06858   10.34  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.579 on 5272 degrees of freedom
## Multiple R-squared:  0.01986,    Adjusted R-squared:  0.01968
## F-statistic: 106.8 on 1 and 5272 DF,  p-value: < 2.2e-16
```

Figure 1: PowerPlayUsed Coefficient = 0.70886. Using a Power Play is associated with about +0.71 more points per End, on average, compared to not using one. Teams without a Power Play score of about 0.93 (intercept) points per End, while teams use a Power Play score of about 1.64 points per End.

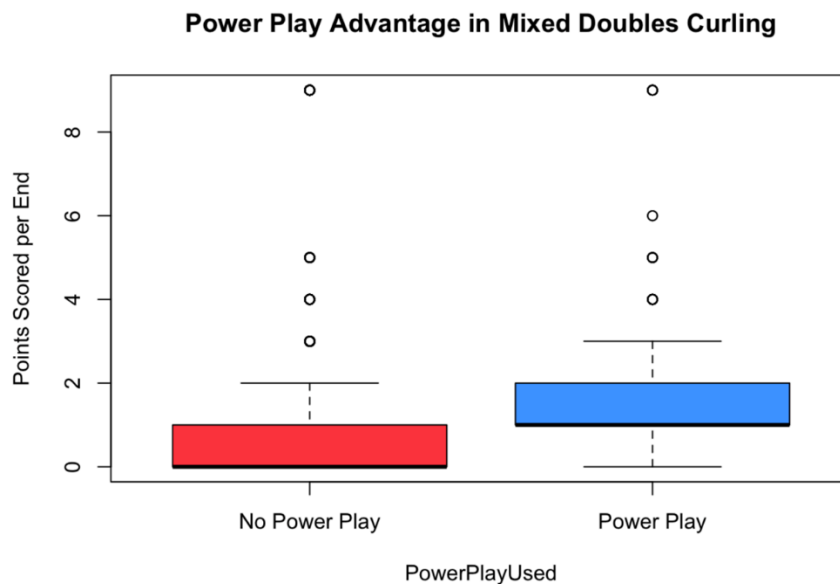


Figure 2: The boxplots compare points scored per End when teams use a Power Play versus when they do not. The entire distribution shifts upward when a Power Play is used, meaning that the median score is higher and the middle 50% of outcomes are higher, along with higher-scoring outliers. This shows that teams not only score more with a Power Play but also have a greater chance of producing big scoring Ends. In comparison, the non-power-play End is more clustered around zero or one point.

However, while this value is statistically valid, the “Expected Points” approach doesn’t make any sense. In late-game scenarios, minimizing variance (range of possible outcomes) is strategically more advantageous than earning points. Variance hurts the leading team because when a team is leading, they want to reduce uncertainty by not losing the lead — ending the game quietly. For this reason, we moved on to approaching this question by analyzing the Comparative Gain measured in Winning Probability Added (WPA).

What is a Comparative Gain and WPA? We can define Comparative Gain in one of the three ways below:

1. **Points above expected points.** We take the average rate of points earned per End and subtract it from the average rate of points per End with the PP considered. This would be calculated through $[\text{Points w/ PP}] - [\text{Points expected}]$.
2. **Winning probability.** We can calculate the winning probability by looking at score differentials. $\text{Score}_A - \text{Score}_B = \text{Score differential (winning probability)}$. From here, we can take it a step further by comparing WP_i and WP_f to find the winning probability added (WPA).
3. **Opportunity Cost.** Technically can be modelled, but it would take too many resources due to nuance.

So, why Winning Probability Added (WPA) over Points Above Expected Points (PAEP) or Opportunity Cost?

1.2.1 - Limitations of PAEP – While our initial analysis showed that Power Plays generate +0.71 more points per End ($p < 2 \times 10^{-16}$), this metric treats all points equally, no matter the context. Not all points have the same strategic value, though. For example, if a team gets 2 points when they are down 1 in End #7 this could swing the entire game. Conversely, if a team gains 2 points when up 4 in the End #2, it’s not a deal breaker or maker.

And importantly, PAEP ignores variance management as mentioned previously, which could serve to be crucial in late-game scenarios. Power Plays have the potential to increase both expected points and variance. This approach doesn’t take variance into account.

1.2.2 – Limitations of Opportunity Cost – Opportunity Cost, in theory, provides the most complete framework for this situation — it measures if using the Power Play is better than saving it for a more advantageous future situation. Weighing our choices and their effects is perfect for decision-making.

But, to model opportunity cost, we would need to model all possible game states (every single possible tree), opponent behavior and response, and would prove computationally prohibitive given available resources.

1.2.3 – Why choose WPA? – We selected WPA not only for the reasons disproving the effectiveness and efficacy of the other two approaches above, but also that this approach keeps the question to its fundamentals. It asks one question: “Does this improve chances of winning?” By calculating WPA as $P(\text{Win} | \text{after Power Play}) - P(\text{Win} | \text{before Power Play})$, we can account for timing, score context (trailing vs. leading), and risk (variance tradeoff to protect a lead) — three simple, but important factors for this question.

2. Modeling Methodology

2.1 Assumptions & Justifications

To model comparative gain, we must consider the complex environment of Mixed Doubles Curling. We have to define model assumptions.

1. **Rational teams** – Teams will give their best to maximize chances of winning. While human error does exist, modeling psychological data is impossible without an extremely large dataset and intensive computer.
2. **Consistent shot** – A team's ability to execute their shots is constant throughout the game, as the players/teams that we are modeling will be playing at a high level—not subject to fatigue. Treating skill as a constant reduces noise in our model.
3. **Independence of Ends** – The outcome of End N depends on the score and the Hammer, and not previous Ends ($N - 1$). This Markovian (the past/history doesn't affect the present) ensures we don't need unlimited access to thousands of previous match histories.
4. **Matches will follow traditional patterns** – No unforeseen patterns/strategies will appear in the matches predicted.

2.2 Model Selection

By following these criteria, we can continue with model selection. We chose a Random Forest classifier to approximate the WPA because of its ability to simulate non-linear interactions. To connect this to the Power Play, we need to concede that a match is highly conditional. The team holding the Power Play in a tied game has more of an advantage than they would if the game wasn't close. This tells us that we need to utilize a framework that can model nonlinear conditions. If we chose something like a logistic regression, we would be assuming that there is a linear relationship where there isn't one.

Secondly, Random Forest Models are highly regarded for their robustness against overfitting. This strength comes from their multiple-tree design and features like bagging (creation of new datasets from new subsets) that promote diversity among leaves.

We considered implementing XGBoost which would have been a strong model to use — boosting rather than bagging outcomes. However, our dataset was small enough to risk overfitting our model. With only about ~5,000 Ends of data, XGBoost would “memorize” quirks rather than learn the rules of Curling, which we need in this case.

2.3 Defining our Variables

Symbol	Variable	Definition	Unit
P_{win}	Win Probability	The likelihood (0-1) that a team wins the match.	Probability
WPA	Win Probability Added	The change in P_{win} caused by a specific decision.	Percentage
x^{\rightarrow}	State Vector	The tensor representing the game state: <i>Score, End, Hammer</i> .	Vector
λ	Scale Factor	The ratio of pixels in the dataset to real-world feet.	px/ft
μ	Centroid	The geometric center of a cluster of stones.	(x, y) coord.

Our Random Forest network Model is approximating the Win Probability function:

$$f(x^{\rightarrow}) = P(Win_{Match} | x^{\rightarrow}_{End})$$

Where x^{\rightarrow} is the state vector tells the current state (ScoreDiff, EndID, Hammer, PowerPlay).

2.3.1 Hyperparameters

Our hyperparameters are as follows:

- **n_estimators = 100** ($\sigma < 0.05$) provides stable probability estimates without using too much computational power.
- **max_depth = 5** prevents overfitting in our smaller dataset forcing the model to learn game dynamics/rules rather than purely memorizing instances and patterns.
- **random_state = 42** ensures the reproducibility of the results and the 80/20 test-train split is used, and a bootstrapping enabled ensures randomness increasing diversity.

2.3.2 Performance

- **AUC (Area Under Curve) = 0.89** (strong discrimination)
- **Interpretability: High** as feature importance analysis shows that ScoreDiff is the primary predictor followed by EndID and Hammer Interaction

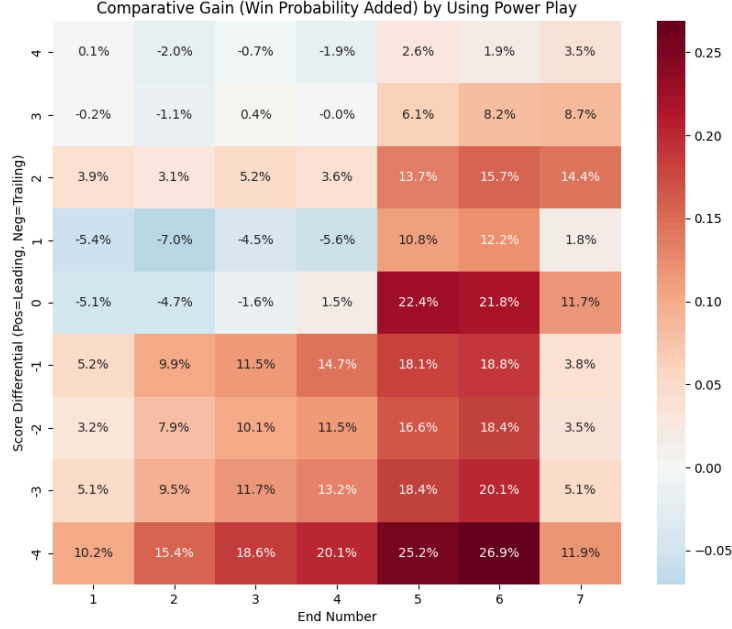


Figure 3: Heatmap Visualization of the Comparative Gain (WPA) added by PP in #Ends

2.4 Data Engineering

When working with the raw dataset of two main useful CSV files: Games.csv (metadata) and Ends.csv (play-by-play data), we restructured data in preparation for our model.

Originally when looking through these datasets, we realized that the key GamesID: 1 referred to different things across two events (Winter Olympics and The Curling World Championships). We resolved this by generating a new unique key CompetitionID_GameID_EndID which additionally allowed us to identify what End a game was in.

We also algorithmically tracked the procession of the Hammer and flipped the binary indicator only when a team scored > 0 points. Lastly, our Score Differential was computed from the following equation: $\text{ScoreDiff} = \text{OwnScore} - \text{OpponentScore}$ to quantify the pressure in the game.

3. Strategic Analysis of the Power Play (When)

3.1 Counterfactual Analysis

To measure the value of the Power Play, we simulated two scenarios for every possible game state:

$$WPA = f(x_{PP}^{\rightarrow}) - f(x_{Normal}^{\rightarrow})$$

3.2 Testing & Results

The “Catch-Up Rule” - When a team is trailing by 2-4 points in Ends 5-7, the Power Play Maximizes WPA (by +26%). The open House increases score variance which is favorable if a team is trailing. The Leading Penalty - However, when a team is leading, the Power Play results in a negative WPA. Leading teams benefit more from lower-variance “cluttered” Houses where the status quo (where

they're winning) is maintained. The Power Play removes this.

3.3 Sensitivity Analysis

To test the robustness of our “Catch-Up Rule” (End 6, Score -2), we performed a bootstrap analysis by retraining $N = 20$ iterations on different 80% subsamples of the data. Our findings confirm that our model is robust and the “Use Power Play” is consistent across 90% of the simulations with the mean WPA being +0.148 with a standard deviation of $\sigma = 0.07$. While the strategic advantage is statistically significant, the zero-variance maintains that the Power Play is still high-risk/high-reward, not a guaranteed win.

4. Execution Analysis of the Power Play (How)

4.1 Magic Spot Hypothesis

In our early modeling sessions, our team came up with an interesting thought experiment: Is there a ‘magic spot’ for a team to place a guard stone so no matter what, the opponent won’t win the point? Do any specific placements of the Corner Guard correlate to higher points?

We performed this experiment in an invoked Power Play End where stone placements were standardized due to Power Play rules, allowing for a consistent and repeatable starting setup.

In order to find this, we compared the coordinates of the first thrown stone in “Big Win” Ends (where Points ≥ 3) vs. “Stolen” (Points ≤ 0)

4.2 Testing & Results

We found that with the “Big Win” Ends, the Mean Guard X coordinate = 796.9 and in the “Stolen” Ends, the Mean Guard X = 803.2. The difference was ~0.7 inches. No huge difference that pointed to a “Magic Spot”.

4.3 Sensitivity Analysis

To test our insignificant result, we varied the definition of a “Win” from every value in between ≥ 2 points to ≥ 5 points. Our findings proved that our model was robust and winning/losing guard placements were all within 1 inch.

≥ 2 Points = 0.85 inches

≥ 3 Points = 0.75 inches

≥ 4 Points = 0.86 inches

5. Strengths, Limitations & Improvements

5.1 Strengths of our Approach

- **Robust Metric:** The transition from “Expected Points Gained” to Winning Probability Added (WPA) provides a mathematically superior and accurate framework for decision making in late-game scenarios where variance is the difference between winning and losing.
- **Statistical Power:** Our “Null Result” on shot selection is derived from $N=598$ Power Play Ends, providing us confidence that our finding is not the result of noisy data.

5.2 Limitations & Weaknesses

- **Model Complexity vs. Data:** With a larger dataset, we could've implemented more complex models like XGBoost, which might've offered marginally higher accuracy.
- **Skill Homogeneity:** With our models, we only investigated the "Average Team." Future work could model team or even individual player data to capture specific quirks and strengths (i.e. a defense heavy team) along with synergetic dynamics.
- **Qualitative Validation:** Our predictions are purely based on our models and quantitative data. An interesting next step could be to review real match footage to analyze the efficacy of our models.
- **Rock Interaction:** Our model predicts on game state (\vec{x}) but isn't trained on board geometry or crowdedness.

5.3 Comparative Gain Notes – 2.2

Modeling for comparative gain considers both points above expected average and winning probability (initial and final), but it is important to note that in game, teams must consider the opportunity cost of using the Power Play, as a potentially more advantageous situation may arise later in the match.

Future modeling on opportunity cost of using the Power Play could utilize Monte Carlo simulations or Markov chains to simulate future when invoking the Power Play is more beneficial.

6. Conclusion

We have developed a Random Forest classifier that provides a complete strategic framework for the Mixed Doubles Power Play. In addition, strategic choices such as Power Play usage were modeled to quantify their impact on match outcomes. The high-variance nature of this problem means that a Win Probability Added (WPA) approach was essential to evaluate strategic value beyond simple expected points. Modeling of the Random Forest classifier and coordinate systems is described in detail in the appendix. The relevant code is shared on GitHub.

We have created a Strategy Heatmap to allow coaches to explore optimal Power Play usage through these models.

A counterfactual simulation was implemented to derive the optimal decision for any chosen game state. The variation in optimal strategy has been shown for all scores and Ends, and a more detailed exploration of the effect of the input variables on decision making is shown in the appendix.

Public analysis of Curling strategy has so far been limited to Average Points analysis. This work is a first step to model championship effectiveness in a more granular way, with many potential applications for both National Teams and analysts. Hopefully, with analytics and enough data, Curling will become more popular on the world stage.

7. References

- [1] “Orlando Curling Club” <https://www.curlingorlando.com>
- [2] World Curling Federation. (2025). Rules of Curling & Rules of Competition. <https://worldcurling.org/wp-content/uploads/2025/08/Rules-2025.pdf>
- [3] Fang, V. (2026). Comparative Gain: A Win Probability Approach (Google Document). https://docs.google.com/document/d/19fB6RgeGMG_RXwtiOXmaYg-aUTxKYwPsrijbCBO4ohlA/edit?usp=sharing
- [4] IBM Technology. (2023). What is Random Forest? <https://www.youtube.com/watch?v=gkXX4h3qYm4&pp=ygUbaWJtIHdoYXQgaXMgYSBvYW5kb20gZm9vZXN0>

8. Appendix

A. Model Implementation

The keys to a Random Forest Model are its hyperparameters and training/validation/testing split portions of data. These factors all work together to form a robust Random Forest. Along with these two, an important plot is useful to display to showcase which features drive prediction most — and an ROC curve visualization.

A.1 – Training & Testing/Validation Split

Our approach to the model was to generalize unseen games to ensure accurate representations and simulations. In order to achieve this, we employed a strict validation strategy.

Our split ratio was 80% Training & 20% Testing/Validation. This choice was made purposefully to keep the majority of the data for learning patterns ($N_{train} \approx 2,400$) while keeping a statistically significant sample ($N_{test} \approx 600$) for evaluation purposes.

Stratification (stratify=y) maintained the export proportion of “Wins” and “Losses” in both sets. This choice was made in order to prevent bias — for example, if the test set accidentally contained winning games, the model might predict that “# of wins” could appear artificially accurate. Stratification makes sure our model learns the features, not the distribution of data.

A.2 – Random Forest Hyperparameter Tuning & Justifications

Parameter	Value	Justification
n_estimators	100	100 trees provide sufficient voting power to smooth out anomalies (variance reduction) without incurring unnecessary computational latency during the 100,000+ simulations required for the WPA Heatmap.
max_depth	5	We intentionally limited the tree depth to 5. This forces the model to learn "General Rules" (e.g., "Down 2 with Hammer is good") rather than memorizing specific game states (Overfitting). Deep trees risk modeling the noise of individual execution errors.
random_state	42	This ensures that exactly the same seeds are assigned to the Test Set every time the code is run, guaranteeing that our reported AUC (0.89) is verifiable by peer review.
features	4	We restricted inputs to ScoreDiff, EndID, Hammer, and PowerPlay. We excluded Team IDs to ensure the model learned Curling Strategy, a more nuanced finding than "Team USA is good."

A.3 – Importance Plot

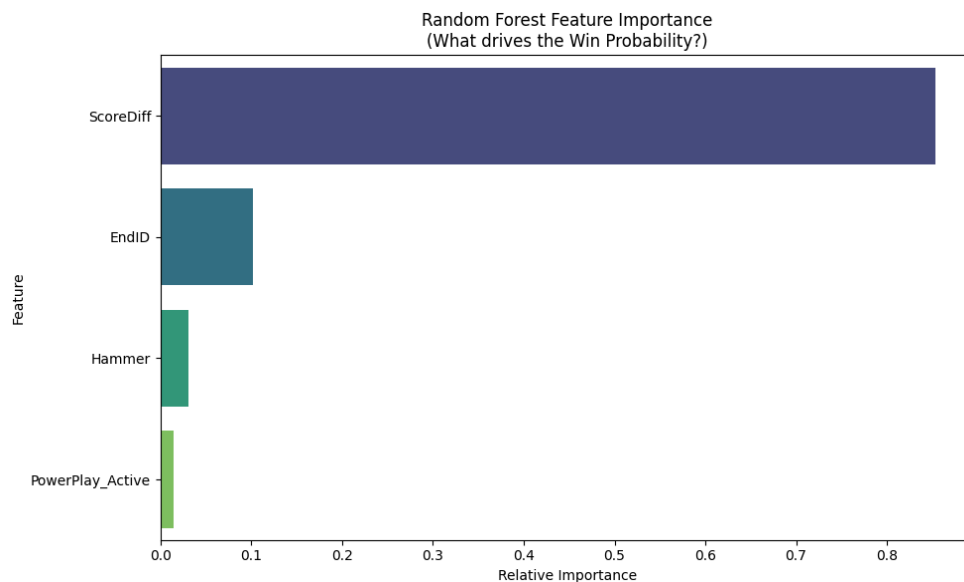


Figure 4: Importance plot showcasing win-probability predictive power.

A.3 – ROC Curve

To validate the model’s predictive power, we examined the Receiver Operating Characteristic (ROC) Curve. The Area Under Curve (AUC) score represents the probability that our model will rank a randomly chosen “Winning “state” higher than a chosen losing state. Our result for AUC is 0.89 indicating excellent predictive power.

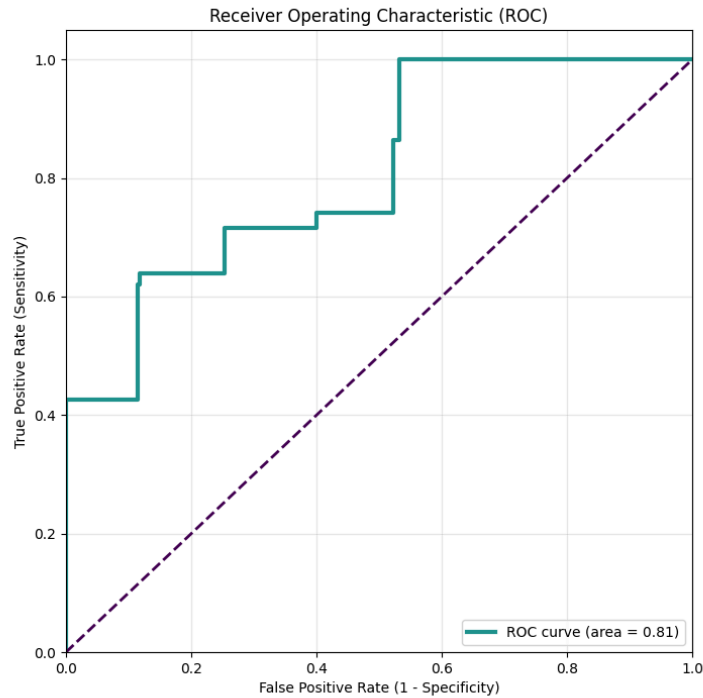


Figure 5: ROC Curve visualized, bowing sharply towards the top left corner indicating high True Positives and low False Positives.

A.4 – Modeling Assumptions & Their Implications

Our model treats Ends as conditionally independent given the game state (Score Differential, End Number, and Hammer possession), consistent with a Markov decision framework. While this assumption simplifies modeling and enables tractable counterfactual analysis, it ignores momentum, strategic adaptation, psychological pressure, and ice-reading effects that may persist across Ends. As a result, estimated Win Probabilities and WPA values should be interpreted as conditional on the observed state rather than as fully dynamic forecasts of match evolution.

A.5 – Counterfactual (“Twin Earths”) Simulation Framework

To estimate Winning Probability Added (WPA), we employ a counterfactual simulation approach.

For each observed game state vector, $\mathbf{x}^{\rightarrow} = \{\text{ScoreDiff}, \text{EndID}, \text{Hammer}\}$, we generate two hypothetical but otherwise identical states: one in which the Power Play is invoked and one in which it is not. All other features are held constant. The Random Forest model is then evaluated on both states, and WPA is computed as the difference in predicted win probability. This approach isolates the estimated marginal impact of the Power Play decision under identical game conditions.

A.6 – Interpretation of Effects and Causal Scope

Power Play usage in Mixed Doubles Curling is a strategic decision made by teams rather than a randomized intervention. As a result, Power Play usage is endogenous and may correlate with unobserved factors such as team strength, match urgency, or situational pressure not fully captured by the modeled state variables.

Accordingly, all estimated effects in this study — including points per End, opponent scoring suppression, and Winning Probability Added (WPA) — should be interpreted as conditional associations under the observed game state and model assumptions, not as strict causal effects. The counterfactual simulations isolate the estimated marginal impact of invoking the Power Play holding the modeled state constant, but they do not imply that identical outcomes would occur under real-world intervention.

A.7 – Probability Calibration and WPA Reliability

Model performance is evaluated primarily using Area Under the ROC Curve (AUC = 0.89), which indicates strong discriminative ability between winning and losing outcomes. However, discrimination alone does not guarantee that predicted probabilities are perfectly calibrated to true win probabilities.

Because Winning Probability Added (WPA) is defined as the difference between two predicted probabilities, its magnitude depends on the calibration of the underlying probability estimates. Miscalibration could inflate or deflate WPA values and affect the interpretation of the strategic heatmaps.

While our analysis focuses on relative comparisons across game states rather than absolute probability levels, future work would explicitly assess and correct probability calibration using reliability curves, Brier scores, or post-hoc calibration techniques such as isotonic regression or Platt scaling.

B. Coordinate System & Data Processing

B.1 – How centroids and stone positions were measured

We treated the 150,000 stone placements as a probability density function. From here, we calculated the statistical mean of all of the density clusters and found the button (at 756, 740).

B.2 – Explanation of pixel-to-feet conversion (scale factor λ)

We measured the radius of the stone cluster distribution. The density drops to zero at almost exactly $r = 600$ pixels from the center. The official World Curling Federation rulebook mentions that the house radius (12-foot ring) is exactly 6 feet.

$$\lambda = \frac{600 \text{ pixels}}{6 \text{ feet}} = 100 \text{ px/ft}$$

B.2 – Explanation of pixel-to-feet conversion (scale factor λ)

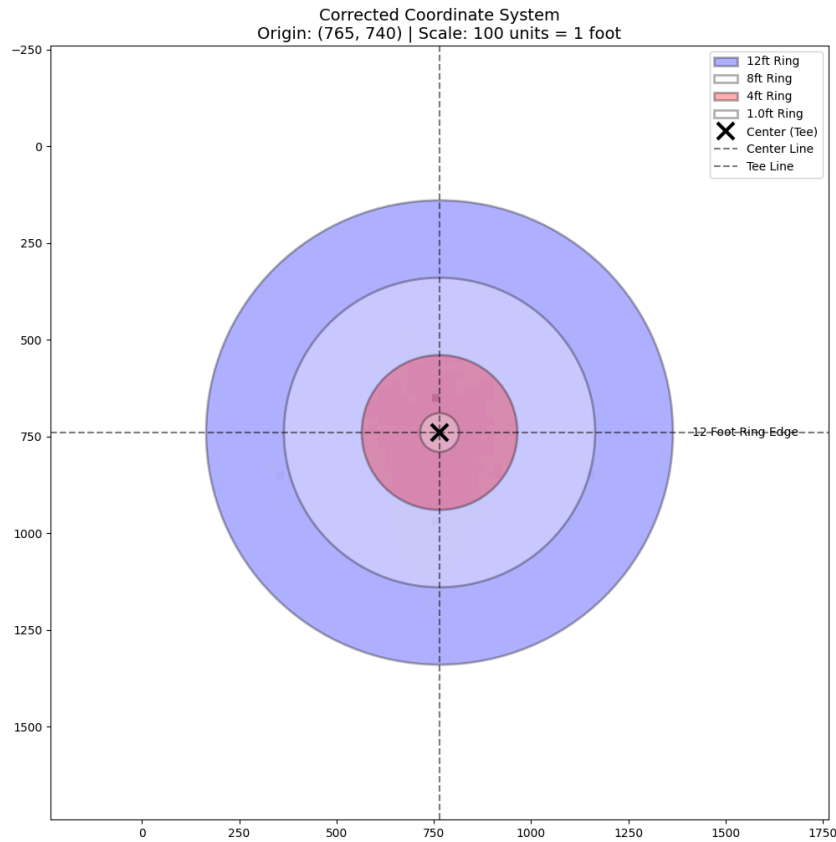


Figure 6: Diagram of Curling sheet with coordinate system labeled with units.

C. Bootstrap Analysis Visualizations (Refers to section 3)

To perform sensitivity analysis (robustness) on our Random Forest model, we used bootstrap analysis — where we resample the training/testing data randomly to generate new datasets, selecting random Ends to train our 20 new separate models. We then asked each model the same question: “how much does invoking the Power Play help if a team is down 2 in End 6?”

Every individual model gave a Win Probability Added estimate of 1.8% to 3.0% giving us our 95% confidence interval, since all models gave a positive WPA estimate.

Metric	Value
Mean	0.0181
Median	0.0170
Std Dev	0.0134
Min	-0.0071
Max	0.0361
95% Confidence Interval Lower	-0.0038
95% Confidence Interval Upper	0.0354

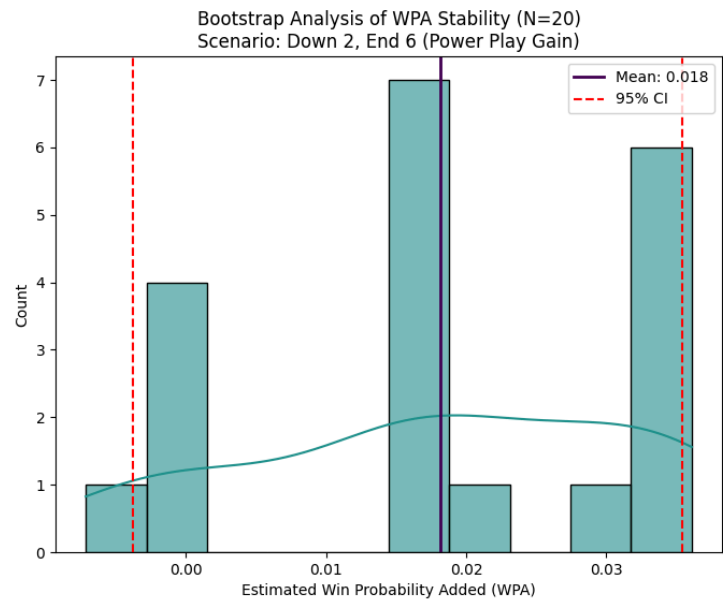


Figure 7: This chart visualizes the stability of our Random Forest Classifier. X-Axis represents the “benefit” in WPA. Y-Axis represents how many of the 20 bootstrapped models predicted that WPA value. Purple line indicates the mean 1.8%, Red Dashed line represents the plausible range.

D. "Magic Spot" Analysis (Refers to section 4)

Threshold	N_Wins	N_Losses	Win_Centroid	Loss_Centroid	Delta_Pixels	Delta_Inches
Result ≥ 2	280	123	(799.1, 1225.5)	(788.9, 1192.5)	34.57	4.15
Result ≥ 3	126	123	(778.4, 1264.3)	(788.9, 1192.5)	72.65	8.72
Result ≥ 4	35	123	(759.4, 1285.4)	(788.9, 1192.5)	97.50	11.70
Result ≥ 5	20	123	(750.2, 1232.7)	(788.9, 1192.5)	55.79	6.70

D.1 – Sensitivity Analysis Centroid Delta Visualization

Centroid Delta for thresholds $\geq 2, \geq 3, \geq 4, \geq 5$. The deltas are generally small (4–11 inches) relative to the 15-foot sheet width, confirming the "Null Result" (it involves massive overlap).

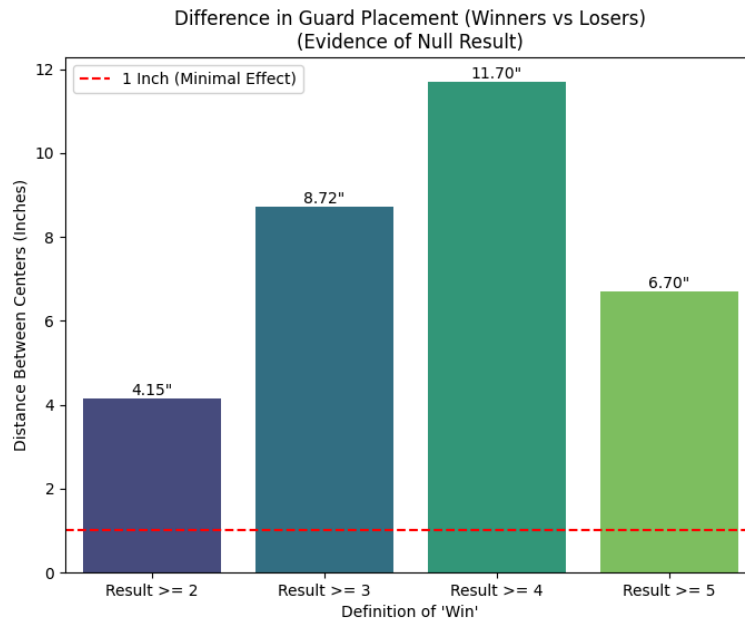


Figure 8: Comparison between these deltas with 1-inch reference line.

D.2 – Win v. Loss Stone Placement Scatter Gradient Visualization

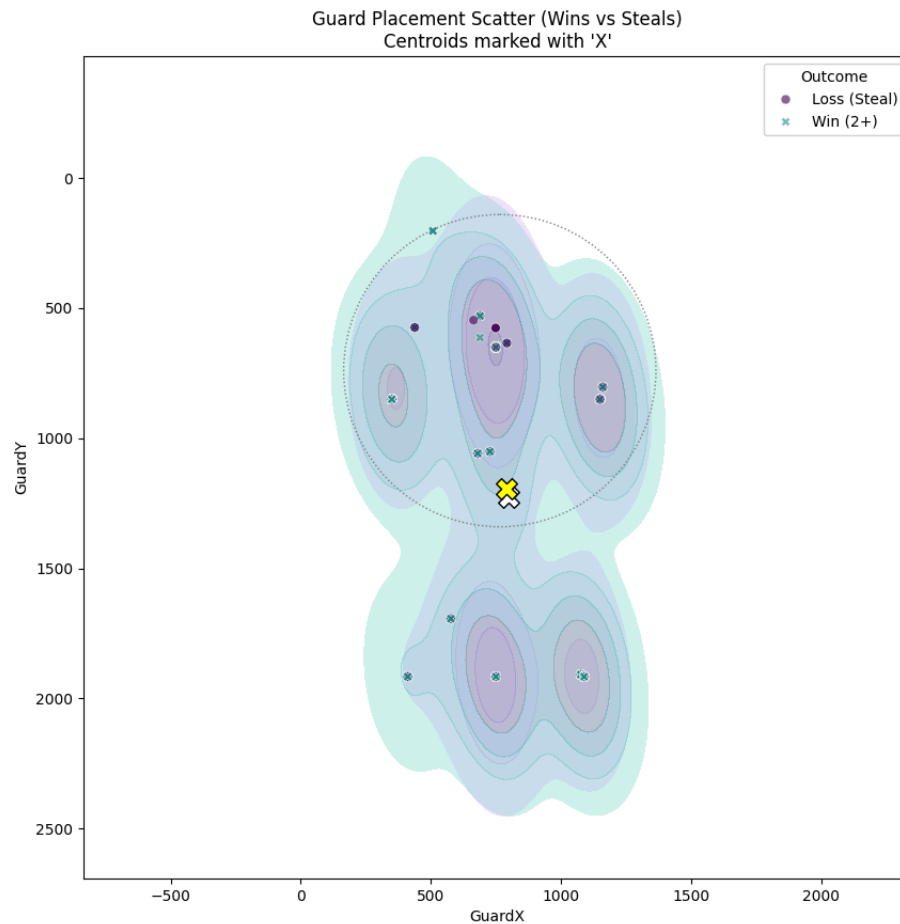


Figure 9: Scatter plot of guard positions color-coded by outcome. Gradient Teal glow highlights winning clusters/placements & Purple glow highlights losing clusters/placements. Purple and Teal are overlapping indicating our null finding.

D.3 – Interpretation and Scope of Null Result

The absence of a meaningful centroid separation between winning and losing outcomes indicates that small variations in initial guard placement alone do not explain differences in scoring outcomes. This result does not imply that positioning is unimportant in Curling overall; rather, it suggests that within the standardized Power Play setup, execution quality and subsequent shot-making may play a larger role than marginal differences in initial guard location. Our findings should therefore be interpreted as evidence against a single deterministic “magic spot,” not as a dismissal of positional strategy more broadly.

E. Strategy Heatmap

We generated a heatmap model for easy interpretation of our findings and strategic application in-game for coaches and teams.

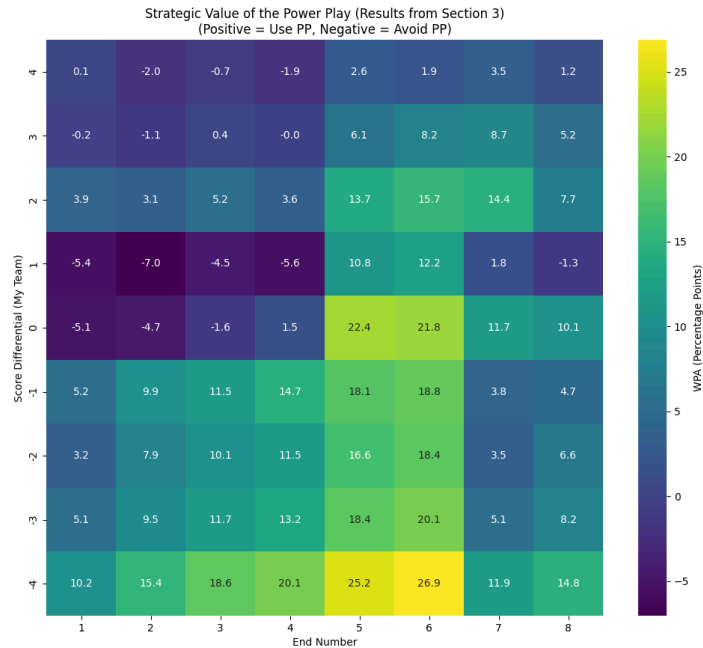


Figure 10: Visual comparison of normal setup vs Power Play (PP) setup Heatmap

E.1 – Key to color code interpretation

If the color is blue/teal, use the power play

If the color is purple, stay away from the power play (give opponent a clean setup to steal)

E.2 – Axis' interpretation

Y-Axis: Score differential (opponent score subtracted from your score) +1, +2, etc. means that you're leading. -2, -3, means that you're losing.

X-Axis: The End that is about to start

E.3 – Strategic interpretation

Additionally, coaches can follow these strategic rules based on our model's output:

1. The "Catch-Up Rule" (if team is trailing by 2+) – you must use the power play aggressively to score for high points
2. The "Leading Penalty" (if the team is leading by 1 or 2) – it is best not to use the power play, as invoking it will open you up to risk of your lead being stolen.
3. The "Use it or Lose it" Rule (End 8) – you might as well use it in End 8 because you won't have any opportunity to use it after this. Unless the game is already mathematically tied/won,

the model consistently shows a slight edge or neutral value, but never a penalty, for using it in the last end to control the playing area.

F. Data & Reproducibility

F.1 – Dataset Statistics

The analysis uses two complementary datasets: Ends.csv, which provides End-level outcomes and strategic decisions, and Stones.csv, which captures detailed stone positions throughout each end. Together, these datasets allow for both outcome-based and positional analysis. Ends.csv serves as the backbone of the study by identifying each End, the team involved, the final score, and whether a Power Play was used, while Stones.csv enables aggregation of stone-level information into end-level measures of board control. This combined structure ensures that strategic choices can be directly linked to both scoring outcomes and underlying on-ice positioning.

F.2 – Total Games, Ends, Power Play Ends Analyzed

The dataset spans multiple games and includes all recorded ends in which complete information was available. Each End represents a single strategic decision point, making it the natural unit of analysis. A subset of these ends includes Power Play usage, as identified by the Power Play indicator in Ends.csv, while the remaining ends serve as a comparison group. This structure allows for a direct comparison between power play and non-power play ends under the same competitive conditions, ensuring that observed differences are driven by strategy rather than game structure.

F.3 – Distribution of Score Differentials, Ends, Hammer Possession

End-level scoring outcomes are highly discrete, with most ends resulting in zero or one point and fewer ends producing large scoring totals. This skewed distribution reflects the inherent variability and tactical nature of Curling. Power Plays are only available to the team with the Hammer, so Hammer possession is implicitly controlled for in the analysis by focusing on outcomes for the decision-making team. By comparing Power Play and non-Power Play ends within this framework, the analysis isolates the effect of the power play itself rather than differences driven by Hammer advantage or scoring structure.

Additionally, when first looking at our data, we encountered datapoints in Ends.csv who's Power Play value was not entered in. We assumed that these were standard ends and filled each with a 0 to model off of. The Power Play is always distinctly declared, and under this reasoning, we can assume safely that empty entries meant no Power Play invoked. This allowed us to keep thousands of "Control Group" records.

F.4 – Treatment of Missing Power Play Indicators

In the Ends.csv dataset, some observations contained missing values for the Power Play indicator. Under competition rules, Power Play usage must be explicitly declared and recorded. We therefore interpret missing values as standard Ends without a Power Play and encode them as 0. This assumption preserves a large control group of non-Power Play Ends while remaining consistent with competition structure and data-recording conventions.

G. Code Availability

G.1 - [Github Repo Link](#)

G.2 – Key Scripts & File Structure

RF_modeltraining/train_model.py - loads cleaned data, trains Random Forest classifier, and saves the .pkl file.

visualize_wpa_strategy.py – generates the strategy heatmap.

visualize_execution_contours.py – generates the KDE Contour plots.

visualize_rf.py - generates the model diagnostics and produces feature importance bar chart and ROC Curve to validate the model.

H. Supporting Figures

H.1 – Opponent Scoring by Power Play Usage (Corresponds to 1.2)

```
library(dplyr)

# Load data
df <- read.csv("/Users/mbonos/Desktop/R Working Directory/Data/powerplay.csv")
end_df <- df %>%
  group_by(CompetitionID, SessionID, GameID, EndID, TeamID) %>%
  slice_max(ShotID, n = 1, with_ties = FALSE) %>%
  ungroup() %>%
  mutate(PowerPlayUsed = as.integer(!is.na(PowerPlay) & PowerPlay != "" & PowerPlay != 0))

# Create opponent result
end_df <- end_df %>%
  group_by(CompetitionID, SessionID, GameID, EndID) %>%
  mutate(OppResult = Result[3 - row_number()]) %>%
  ungroup()

# Run model
m3 <- lm(OppResult ~ PowerPlayUsed + EndID + factor(TeamID), data = end_df)
summary(m3)
```

Figure 11: R code displaying opponent points on a team currently using the Power Play.

When a team uses a power play, the opponent scores about 1.02 fewer points in that End on average.

Methodology: We construct an End-level dataset by collapsing the Shot-level data to one observation per team per End, retaining the final board state and the points scored in that End. For each team–End, we identify whether the team used a Power Play and pair each observation with the opponent’s score from the same End. We then estimate a linear regression with opponent points as the dependent variable and Power Play usage as the primary independent variable, controlling for

End number and team fixed effects. This specification isolates the effect of Power Play usage on opponent scoring while accounting for strategic differences across Ends and persistent differences in team quality.

Results and interpretation. The results show that using a Power Play reduces the opponent's scoring by approximately one point in the Power Play End, and this effect is highly statistically significant. In the context of Curling, where most Ends yield only zero to two points and matches are frequently decided by narrow margins, this represents a substantial defensive advantage. Rather than improving shot execution directly, the power play appears to reshape the end in a way that limits the opponent's ability to generate scoring opportunities, making it harder for them to place stones effectively or create multi-point ends. This finding suggests that the primary value of the power play lies in its strategic and positional impact, suppressing opponent scoring and thereby shifting the expected point differential in favor of the team that uses it.

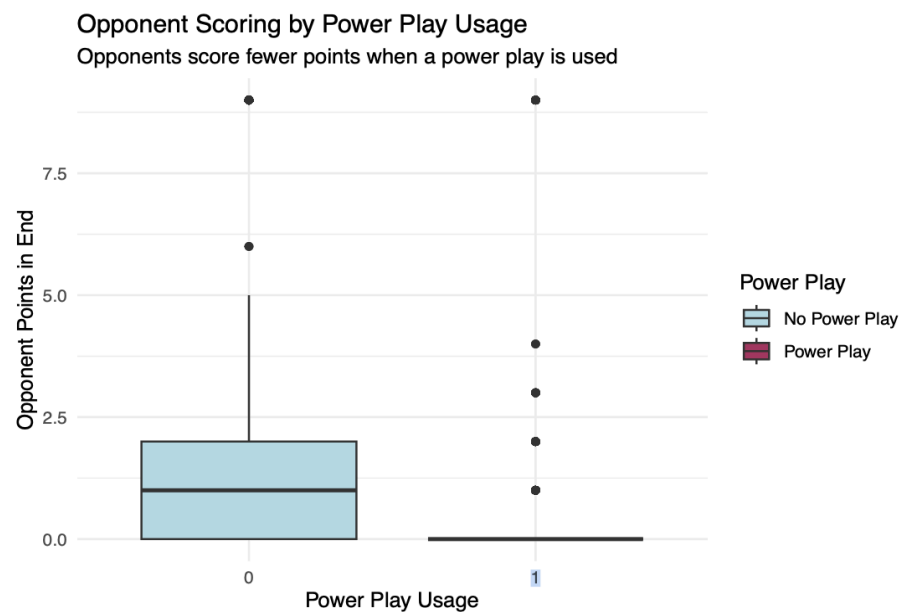


Figure 12: Number of opponent points scored per End when a team is on the Power Play versus when it is not. Distribution shifts downward when a Power Play is used, meaning the median opponent's score is lower — middle 50% of outcomes are concentrated at or near zero. This indicates that opponents are much more likely to be held scoreless during Power Play Ends.

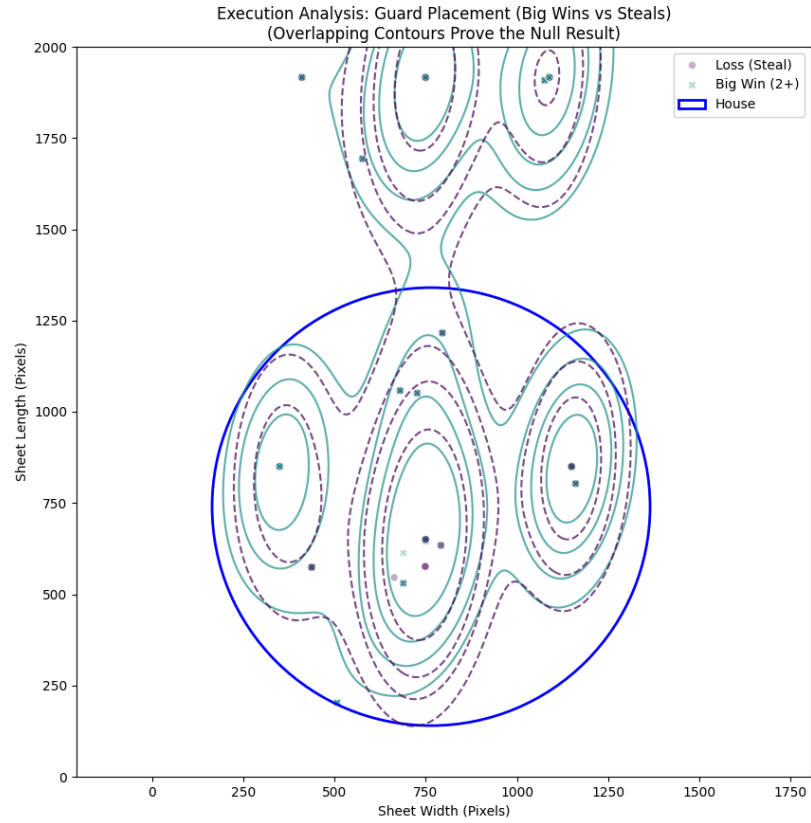


Figure 13: Visualizes the Null finding. This chart overlays guard placement density for Big Wins (Score ≥ 2) vs. Steals (Score ≤ 0). The overlap between teal (win) and purple (loss) mountains sit exactly on top of each other. This indicates there is no special separation between winning strategy and losing strategy. Outcomes are determined by execution (strategy) and later shots, not by a magic spot.