Evaluating Matchmaking Fairness and Winning in Dota 2 by Bayesian: Skill or Chance?

Yuexin (Joy) Wang

ABSTRACT

As esports continues to expand, the integrity of matchmaking systems becomes increasingly important for both player experience and competitive fairness. This study evaluates the fairness of Dota 2's matchmaking by analyzing over 50,000 ranked matches using Bayesian hierarchical modeling. We estimate latent player skill levels and investigate how factors such as hero selection, team composition, and in-game performance metrics influence match outcomes. Our goal is to determine whether observed win/loss patterns align with expectations under a fair system, or if systematic imbalances exist. The findings aim to inform both developers—seeking to improve matchmaking algorithms—and players—looking to optimize strategies for consistent performance.

INTRODUCTION

As multiplayer online battle arena (MOBA) games like Dota 2 become increasingly competitive and data-rich, concerns around the fairness of matchmaking systems are growing. While these systems aim to form balanced teams based on player skill, personal gameplay often suggests otherwise—streaks of wins or losses can feel random, unearned, or even rigged. This project investigates whether such patterns stem from flaws in Dota 2's matchmaking algorithm or are simply the result of inherent randomness. Using Bayesian hierarchical modeling and player-level match data, we aim to assess matchmaking fairness and identify key factors that influence game outcomes.

BACKGROUND

Dota 2 is a popular multiplayer online battle arena (MOBA) game where two teams of five players compete to destroy the opposing team's base, known as the Ancient. Each player selects a unique hero from a pool of over 100, with distinct roles such as carry, support, or initiator. The game involves complex strategies, including hero drafting, resource control (like gold and experience), team coordination, and objective timing.

Matches are shaped by a wide range of variables—from individual player skill and hero selection to in-game economy and real-time decisions. While the game uses a matchmaking rating (MMR) system to balance teams based on past performance, players often experience win/loss streaks that feel disconnected from how well they actually played. This raises questions about whether the system consistently produces fair matches, or if randomness and systemic biases play a larger role.

To explore this, we analyze a dataset of around 50,000 public matches, focusing on players with identifiable accounts and sufficient match history. We apply Bayesian hierarchical modeling to estimate hidden player skill and evaluate how various in-game factors—such as hero choice, team composition, damage dealt, and gold earned—influence match outcomes. Our goal is to assess the fairness of matchmaking and identify structural patterns that may impact competitive balance. These insights could support improvements to matchmaking systems and inform players seeking more consistent results.

DATA

0.1 Source

The dataset used in this study was sourced from Kaggle: "Dota 2 Matches" by Devin Anzelmo, available at https://www.kaggle.com/datasets/devinanzelmo/dota-2-matches. It contains structured match data collected through OpenDota, a third-party analytics platform that accesses Dota 2's public match history via Valve's API.

This dataset includes detailed statistics from over 50,000 ranked matches. Each match record provides metadata such as match ID, duration, game mode, and the winning team, along with in-depth player-level statistics including

hero selection, kills, deaths, assists, last hits, gold per minute, experience per minute, hero damage, and tower damage. Team-level data such as total gold and XP advantage, objective control, and drafting order are also available.

For this project, we limited our scope to matches played in Captains Mode, a format commonly used in ranked and professional games where players draft heroes in a structured, strategic order. This ensured that the analysis focused on competitively meaningful matches with intentional team compositions.

It is important to note that the dataset is static and does not reflect ongoing updates or changes to the game. The data represents a snapshot of gameplay and matchmaking dynamics at the time of collection, and may not align with current game patches or balance changes.

0.2 Pre-processing

To prepare the data for analysis, we applied several filtering steps to improve consistency and ensure the reliability of the results. First, we limited the dataset to matches played in Captains Mode, a competitive game format in which players take turns drafting heroes for their team. This mode is used in ranked matchmaking and professional tournaments, making it ideal for studying strategic team composition and balance.

Second, we excluded all matches that included anonymous or unidentifiable players. Since anonymous players do not have persistent account IDs, their inclusion would make it impossible to track individual performance over multiple games, which is essential for estimating latent skill. We also removed incomplete records with missing or invalid values, such as zero-duration games or null hero assignments.

After these filters were applied, the dataset was reduced from over 50,000 to approximately 4,774 matches, containing only valid, complete records. The final dataset supports two levels of variable structure: match-level and player-level. Our analysis focuses primarily on the player level, where each data point represents one player in a specific match. At this level, we collected variables such as hero ID, kills, deaths, assists, gold earned, damage dealt, and experience gained. Match-level variables, such as match duration, game outcome, and team compositions, were included to provide contextual information and to support model hierarchies. By narrowing the scope to high-quality, consistent match data and isolating individual player behavior, we ensured that our Bayesian models could make meaningful inferences about skill, performance, and fairness in the matchmaking system.

0.3 Pre-analysis

Before applying Bayesian modeling, we conducted an exploratory analysis to understand basic patterns in the data and assess whether matchmaking outcomes showed signs of fairness or imbalance.

The first visualization shows the distribution of win rates among players who have played more than 10 matches. The distribution is tightly clustered around 48% to 50%, with most players falling within a relatively narrow band. This suggests that, over time, the matchmaking system tends to balance win rates, indicating a degree of fairness for experienced players.

In contrast, the second plot focuses on players who have played between 3 and 10 games. This distribution is noticeably wider, with more players having either very high or very low win rates. This spread suggests that newer players experience more variability in their outcomes, likely due to a combination of matchmaking volatility and unstable skill estimation in early stages.

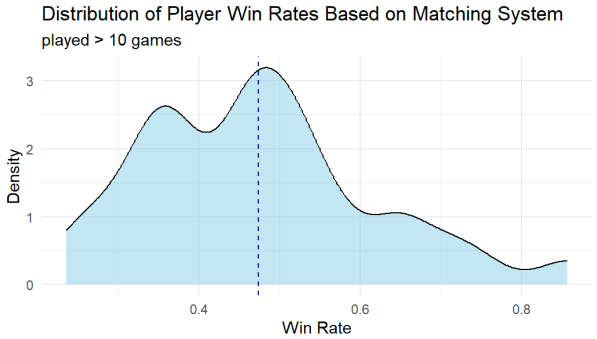


Figure 1. Win Rate Distribution (>10 Games, n = 1246)

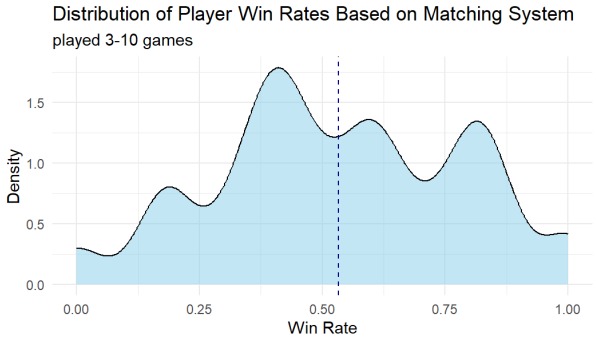


Figure 2. Win Rate Distribution (3-10 Games, n = 450)

The third plot compares average win rates across different experience levels, grouped by number of games played. It shows a clear convergence: players with more games tend to approach the 50% mark, while those with fewer games show larger deviations. This reinforces the patterns observed in the first two plots and highlights a trend of increasing fairness as more data is accumulated for each player.

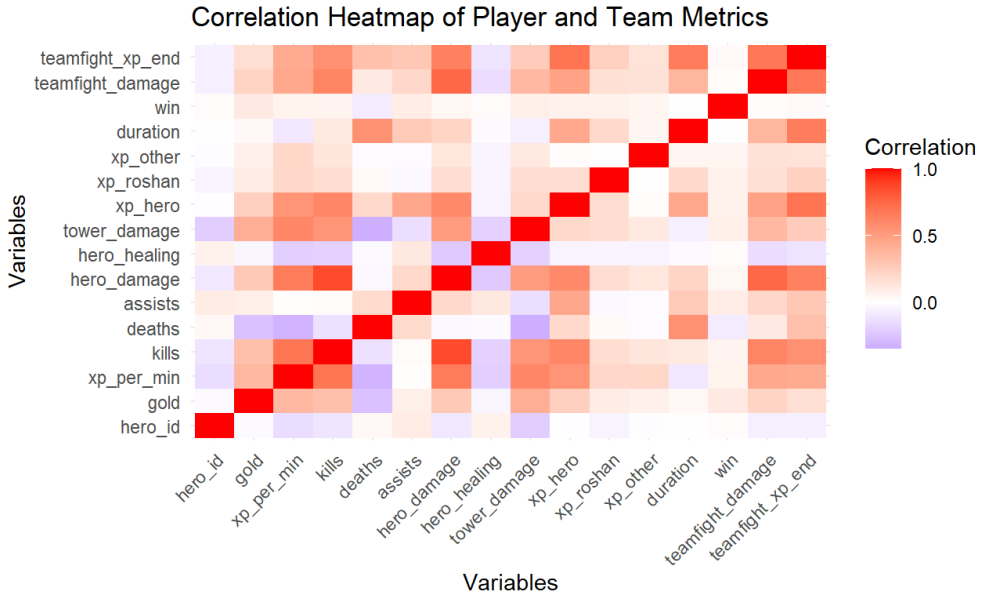


Figure 3. Correlation Heatmap

Taken together, the three visualizations suggest that while Dota 2's matchmaking system may be volatile for newer players, it gradually achieves balance as players continue to participate in more games. These observations support our decision to focus on players with at least 10 matches for the main analysis, ensuring a more stable foundation for estimating player skill and evaluating system fairness.

MODEL

To dig deeper into our research interests, three Bayesian models are utilized. First, a Beta-Binomial Hierarchical Model is constructed to investigate the fairness of winning the game for different experienced players under Dota2's matching system. Next, two logistic regression models were developed to estimate the key factors contributing to match outcomes: one focusing on individual player performance, and the other on team-level performance.

1.1 Fairness of Winning Games: Inexperienced and Experienced Players

To investigate the fairness of Dota2's matchmaking system, we divide players into two categories based on their total number of games played:

- **Inexperienced players**: those who have played between 3 to 10 games
- **Experienced players**: those who have played more than 10 games

Using the beta-binomial hierarchical model, our objective is to evaluate whether the game matchmaking mechanism provides a fair chance of winning, that is, a probability of win of approximately 50% for both groups, separately.

Define two groups of players: the inexperienced group and the experienced group, indexed by $e \in \{1,2\}$. Let $y_{i|e}$ the number of wins for player i in group e, $N_{i|e}$ the number of games played by player i, and $\theta_{i|e}$ is the individual winning probability of player i in group e.

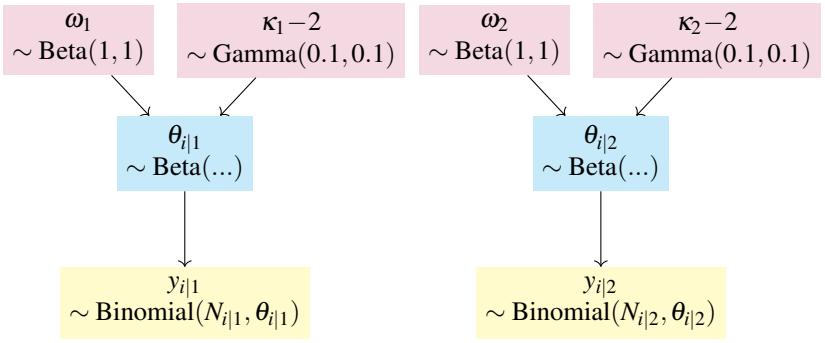
$$y_{i|e} \sim Binomial(N_{i|e}, \theta_{i|e})$$

The model assumes that each player's win rate $\theta_{i|e}$ follows a Beta distribution with group-specific parameters ω_e and κ_e:

$$\theta_{i|e} \sim Beta(\omega_e(\kappa_e - 2) + 1, (1 - \omega_e)(\kappa_e - 2) + 1)$$

We set hyperpriors on the group-level parameters with weakly informative distributions:

$$\omega_e \sim Beta(1,1) \qquad \kappa_e - 2 \sim Gamma(0.1, 0.1)$$



1.2 Individual-Level Logistic Regression: Player Performance and Win Probability

To better understand how individual performance in the game contributes to match outcomes, we developed a Bayesian logistic regression model at the player level. Each observation represents a single player in a match, and the binary outcome variable indicates whether the player's team won. The goal of this model is to estimate how various performance factors affect the probability of victory from the perspective of individual contributions. The outcome variable is binary. It equals $y_i = 1$ if player i's team won, and $y_i = 0$ if they lost.

$$\log\left(\frac{P(y_i = 1)}{1 - P(y_i = 1)}\right) = \theta_0 + \theta_1 x_{1i} + \theta_2 x_{2i} + \theta_3 x_{3i} + \theta_4 x_{4i} + \theta_5 x_{5i} + \theta_6 x_{6i} + \theta_7 x_{7i}$$

- θ_0: Intercept term
- θ_1: Effect of **kills**
- θ_2: Effect of **deaths**
- θ_3: Effect of **assists**
- θ_4: Effect of **gold**
- θ_5: Effect of **teamfight damage**
- θ_6: Effect of **teamfight experience**
- θ_7: Effect of **duration**

The predictors included in the model are standardized values of player kills, deaths, assists, gold earned, teamfight damage dealt, teamfight experience gained, and the duration of the match. These features were selected to capture both offensive and strategic aspects of a player's game play. The logistic regression equation models the log-odds of winning based on these variables, enabling us to interpret the coefficients as the direction and strength of association between each feature and the chance of winning.

Specifically, this model uses non-informative priors for all parameters, as no explicit prior distributions were specified in the Stan code. This design choice allows the data to entirely determine the posterior estimates without introducing subjective bias from prior beliefs. While this approach is often acceptable when large amounts of data are available, it may reduce regularization and introduce more uncertainty in parameter estimates.

1.3 Group-Level Logistic Regression: Performance Differentials and Match Outcomes

To complement our individual-level analysis, we constructed a group-level logistic regression model to examine how team-level performance differentials influence match outcomes. Specifically, we aggregated individual player statistics (kills, gold, teamfight damage, and teamfight experience) to the team level and computed the differences

between the Radiant and Dire teams for each match. These differences (Radiant minus Dire) effectively capture the relative advantages of one team over the other. For example, kill differentials serve as a concise indicator for offensive performance and are inherently related to both assists and the death of the opposing team.

The preprocessing steps included grouping players into their respective teams based on their player slot, summing key performance metrics by match and team, calculating between-team differences, and retaining only Radiant entries with a binary win indicator (1 if Radiant won, 0 otherwise). All difference variables were standardized along with the duration of the match, which was also scaled independently.

$$\log\left(\frac{P(\text{win}_i = 1)}{1 - P(\text{win}_i = 1)}\right) = \alpha + \theta_1 \cdot \text{diff\_kills}_i + \theta_2 \cdot \text{diff\_gold}_i + \theta_3 \cdot \text{diff\_teamfight\_dmg}_i + \theta_4 \cdot \text{diff\_teamfight\_xp}_i + \theta_5 \cdot \text{duration}_i$$

- α: Intercept
- θ_1: Effect of **kill difference**
- θ_2: Effect of **gold difference**
- θ_3: Effect of **teamfight damage difference**
- θ_4: Effect of **teamfight experience difference**
- θ_5: Effect of **match duration**

In this group-level model, we applied weakly informative priors to the coefficients: $\theta \sim \mathcal{N}(0, 0.5)$ and $\alpha \sim \mathcal{N}(0, 1)$. These priors reflect the belief that most effects are likely centered near zero, but moderate deviations are still plausible. This stands in contrast to the individual-level model, in which we used non-informative priors. We chose weakly informative priors at the group level for two main reasons. First, the posterior distributions of the individual-level model were already approximately normal and centered near zero, suggesting that a mild regularizing prior would not distort the results. Second, group-level data are aggregated and typically more stable, but still vulnerable to multicollinearity or overfitting in logistic models. Introducing weak priors helps mitigate these issues and improves the model's generalizability and predictive performance.

RESULTS

2.1 Fairness of Winning Games: Inexperienced and Experienced Players

The following results are from the Markov Chain Monte Carlo (MCMC) model implemented in **Stan** using 2 chains with 10,000 iterations and 1,000 warm-up steps each. To estimate each group's underlying winning rate, we compute the expectation using the formula below applying the sampled values of ω and κ for each group.

$$\mathbb{E}[\theta_e \mid \text{data}] = \mathbb{E}_{\omega_e, \kappa_e \mid \text{data}}\left[\mathbb{E}[\theta_e \mid \omega_e, \kappa_e]\right] = \mathbb{E}_{\omega_e, \kappa_e \mid \text{data}}\left[\frac{\alpha_e}{\alpha_e + \beta_e}\right] = \mathbb{E}_{\omega_e, \kappa_e \mid \text{data}}\left[\frac{\omega_e(\kappa_e - 2) + 1}{\kappa_e}\right]$$

where $e \in \{1, 2\}$, 1 = inexperienced players and 2 = experienced players.

For both inexperienced and experienced player groups, the trace plots of the posterior ω and κ (Figure 8 & 9) show good convergence. The scatter plots (Figure 10 & 11) suggest no strong autocorrelation between parameter pairs, with clear dispersion across the parameter space. Additionally, the effective sample sizes for ω and κ (Table 1) are all above 1,000, indicating reliable sampling efficiency.

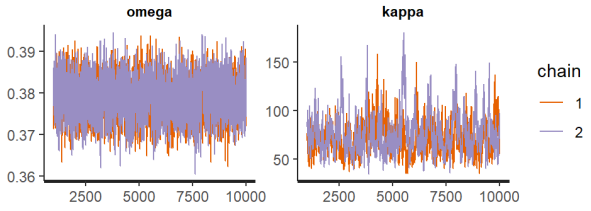


Figure 4. Trace Plots: Inexperienced Players

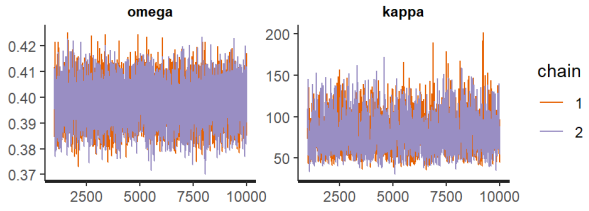


Figure 5. Trace Plots: Experienced Players

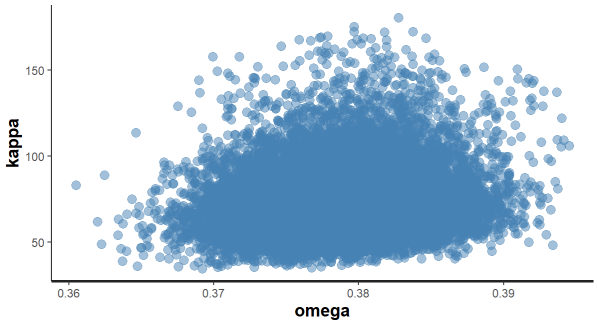


Figure 6. Autocorrelation: Inexperienced Players

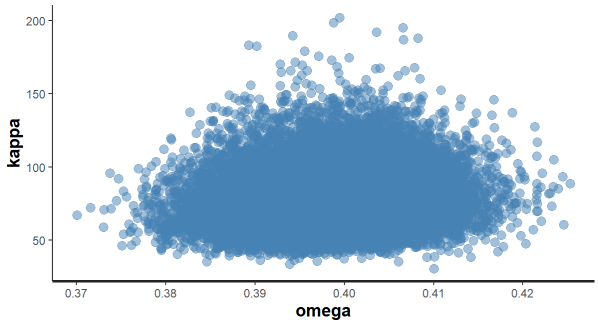


Figure 7. Autocorrelation: Experienced Players

	Inexperienced Players		**Experienced Players**	
	ω_1	$\kappa_1 - 2$	ω_2	$\kappa_2 - 2$
ESS	2323	1721	3527	1135

Table 1. Effective Sample Size of Posterior Group Winning Rate

Then we analyze the posterior distributions of the expected group-level winning rates, $E(\theta_e)$. For both inexperienced and experienced players, the posterior means are below 0.5, and the 95% credible intervals exclude 0.5 (Figure 12 & 13, Table 2). These results suggest that the matchmaking system is systematically biased, favoring losses over wins across both player categories.

	Inexperienced Players	**Experienced Players**
Mean	0.382	0.401
95% Credible Interval	(0.37, 0.39)	(0.39, 0.41)

Table 2. Statistics of Posterior Group Winning Rate

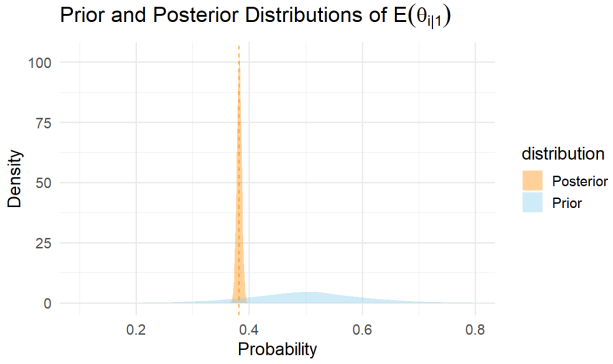


Figure 8. Distribution of group winning rate: Inexperienced Players

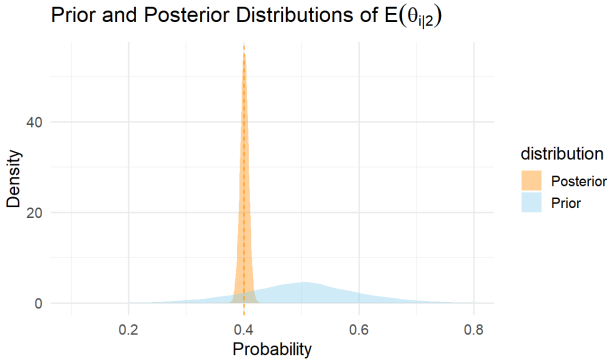


Figure 9. Distribution of group winning rate: Experienced Players

2.1.1 Prior Sensitivity Analysis

We assess prior sensitivity by modifying the prior distributions for ω_e and κ_e. Specifically, we change the distribution of ω_e from $Beta(1,1)$ to $Beta(3,1)$ and the distribution of $\kappa_e - 2$ from $Gamma(0.1,0.1)$ to $Gamma(0.1,1.1)$. The resulting posterior distributions for the two expected group-level winning rates (Figure 10 & 11), exhibit minimal

differences compared to the original setup (Figure 8 & 9). This indicates that our inference is robust to reasonable changes in prior assumptions.

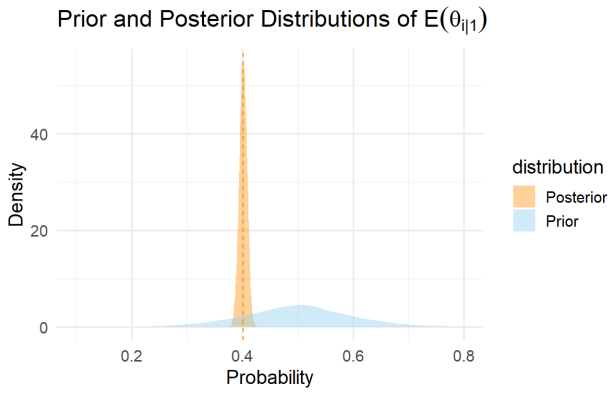


Figure 10. Distribution of group winning rate: Inexperienced Players *(Prior Sensitivity Analysis)*

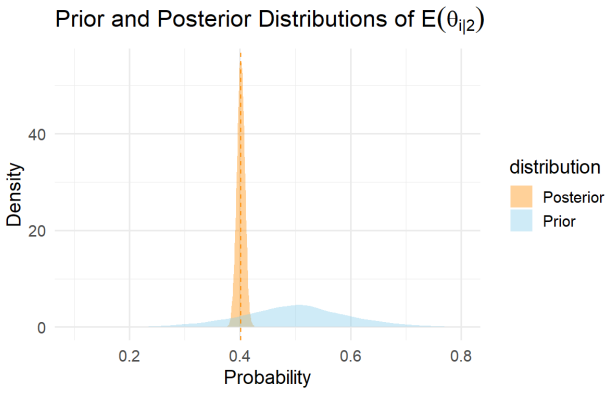


Figure 11. Distribution of group winning rate: Experienced Players *(Prior Sensitivity Analysis)*

2.2 Determinants of Match Outcomes: Individual and Group-Level Effects

We run both the logistics regression models using Bayesian inference in STAN. The setup included 2 chains, 10,000 iterations, and 1,000 warmup steps. The posterior results showed good convergence and low autocorrelation. All parameters had effective sample sizes greater than 10,000.

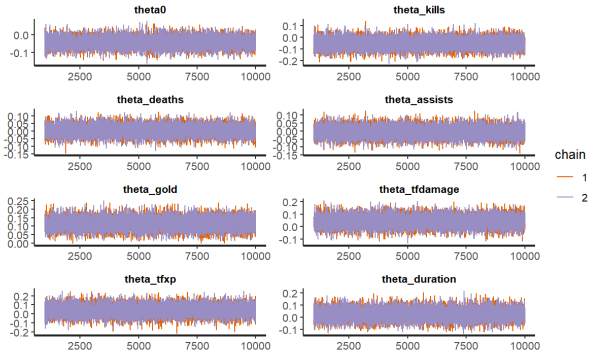


Figure 12. Trace Plots: Individual Player-Level

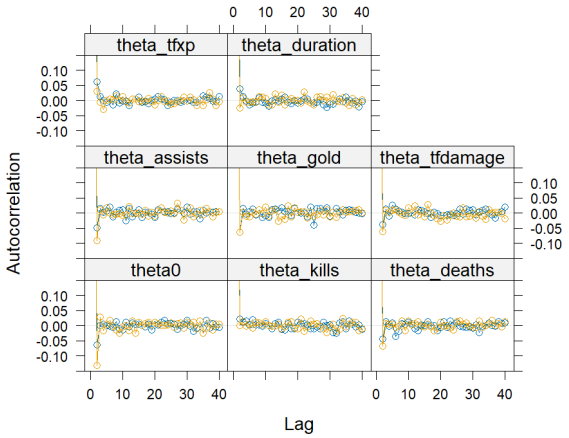


Figure 13. Autocorrelation: Individual Level

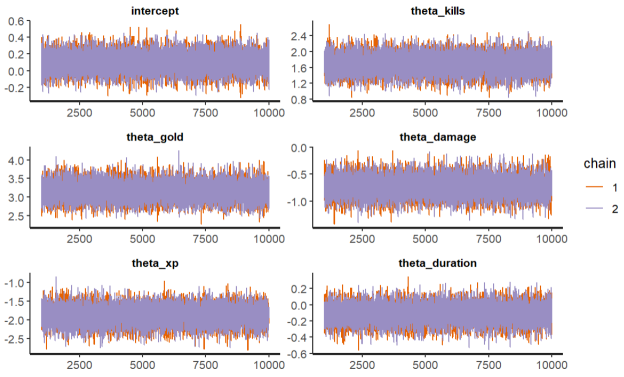


Figure 14. Trace Plots: Group-Level

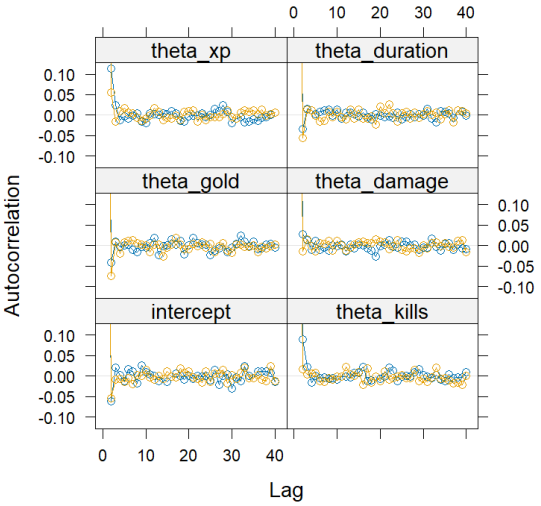


Figure 15. Autocorrelation: Group Level

	$\theta_{intercept}$	θ_{assists}	θ_{deaths}	θ_{duration}	θ_{gold}	θ_{kills}	$\theta_{\text{team fight damage}}$	$\theta_{\text{team fight xp}}$
ESS	12445.4	10736.5	10923.0	17769.5	10439.1	17620.5	10111.7	17024.3

Table 3. Effective Sample Size of Posterior Parameters: Individual Player Level

	$\theta_{intercept}$	θ_{damage}	θ_{duration}	θ_{gold}	θ_{kills}	θ_{xp}
ESS	10331.1	17520.5	10671.5	10212.2	16298.3	15527.3

Table 4. Effective Sample Size of Posterior Parameters: Group Level

From our individual player-level logistic regression model, we can see that gold accumulation has the strongest positive effect on outcomes, with its 95% credible interval entirely above zero. In contrast, kills and assists show weak negative associations, whereas deaths have a negligible effect centered around zero. The remaining variables, teamfight damage, teamfight experience, and duration, have modest positive effects, although some of their credible intervals include zero. Overall, individual performance indicators show limited influence on the final outcome.

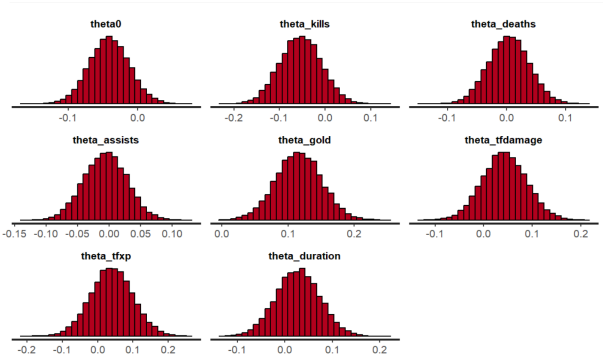


Figure 16. Posterior Distributions: Individual Parameters

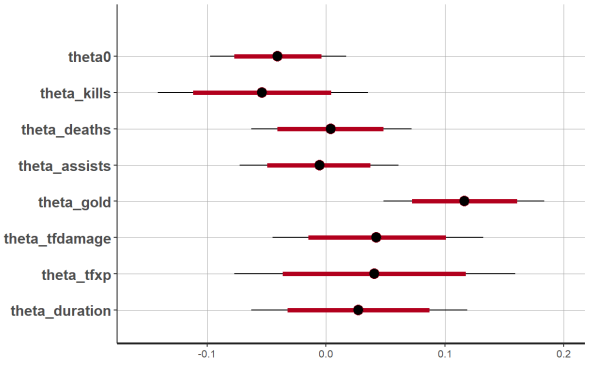


Figure 17. Posterior Statistical Summary: Individual Level

From our group-level logistic regression model, we find that gold differential between Radiant and Dire has the strongest positive effect on the probability of winning, with a 95% credible interval of [2.70, 3.63]. This suggests that, at the aggregate level, outperforming the opposing team economically is a highly reliable predictor of success. Kill differential also has a substantial positive effect (95% CI: [1.23, 2.13]), consistent with the intuition that superior offensive performance benefits the team. Interestingly, both teamfight damage and experience differentials have negative coefficients (95% CIs: [-1.09, -0.38] and [-2.36, -1.42], respectively). This might reflect cases where aggressive or risky engagements yielded high raw numbers without translating into victory, possibly due to poor coordination or inefficient trades.

Match duration, however, shows a small and statistically non-significant effect (95% CI: [-0.33, 0.11]), suggesting limited predictive value.

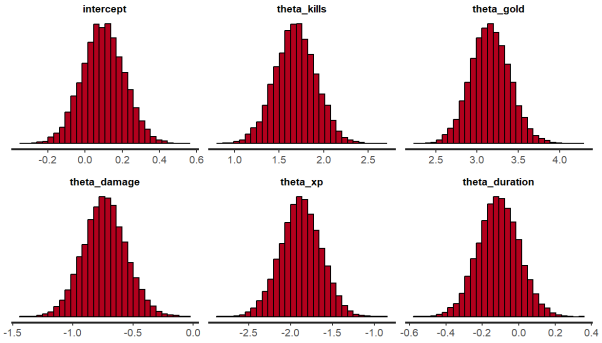


Figure 18. Posterior Distributions: Group Parameters

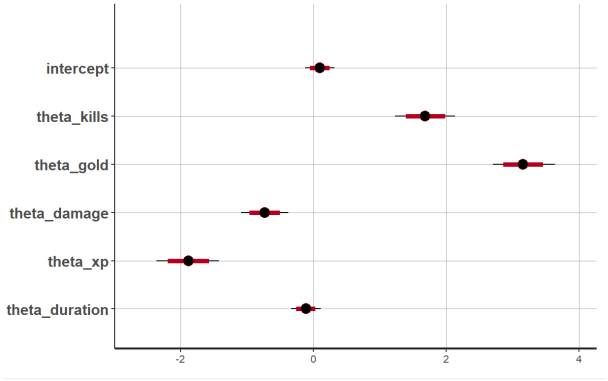


Figure 19. Posterior Statistical Summary: Group Level

2.2.2 Prior Sensitivity Analysis

In the individual player-level logistic regression model, we initially apply implicit non-informative priors by omitting explicit prior specification on the coefficients. To assess prior sensitivity, we re-estimate the model with alternative prior setting: weakly informative priors $\theta_{\mathbf{j}} \sim \mathbf{Normal}(\mathbf{0}, \mathbf{1})$.

After re-fitting the model under these alternative priors, we find that the posterior means and 95% credible intervals of key coefficients (kills, deaths, assists, gold, teamfight damage, teamfight experience, and duration) (Figure 18) exhibit minimal changes compared to the original estimates. Posterior densities largely overlap across different prior choices.

For the team-level logistic regression, we initially place moderately informative priors: $Normal(0,1)$ for intercept, $Normal(0,0.5)$ for other coefficients (e.g., differences in kills, gold, teamfight damage, teamfight xp (experience), and duration). To examine the sensitivity to prior choice, we re-estimate the model under slightly broader priors: $\mathbf{Normal}(\mathbf{0},\mathbf{2})$ for all coefficients.

The comparison shows that posterior estimates, 95% credible intervals of key predictors remained stable (Figure 19). No substantial shifts in the posterior means or uncertainty intervals are observed. These results suggest that the model's posterior inferences are robust to reasonable changes in prior assumptions at both the individual player and group levels.

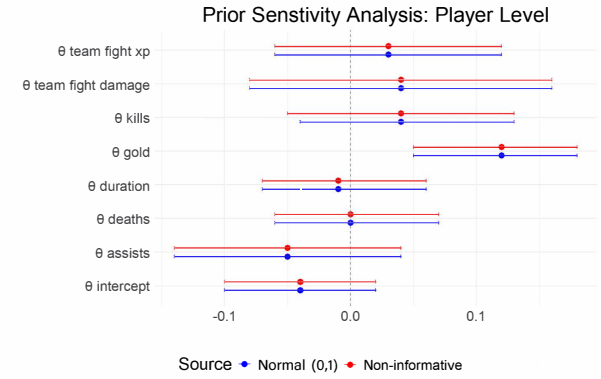


Figure 20. Prior Sensitivity Analysis: Player Level

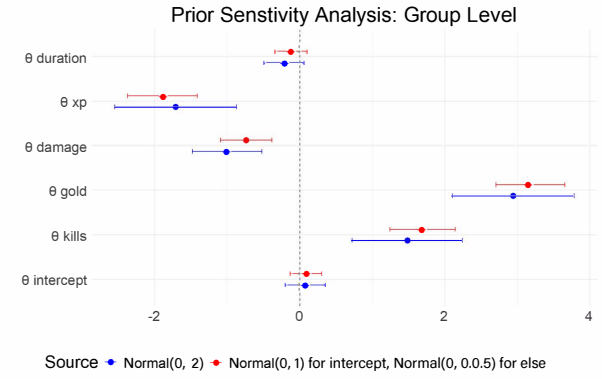


Figure 21. Prior Sensitivity Analysis: Group Level

Thus, these findings underscore a fundamental truth in competitive team games: while individual skill contributes to performance, sustained success is achieved through collective coordination and strategic unity.

DISCUSSION, LIMITATIONS & FUTURE DIRECTIONS

The findings from our model suggest that Dota 2's matchmaking system may exhibit some degree of unfairness, particularly when considering individual performance alone. While individual metrics such as kills, deaths, and assists are important, they appear to have a limited influence on the overall match outcome. Success in the game is largely driven by team coordination and collective effort, rather than the performance of a single player. Even if one player performs exceptionally well, the game's outcome is still heavily dependent on the team as a whole, highlighting the importance of teamwork in multiplayer online battle arena (MOBA) games like Dota 2.

Several limitations exist in this study. First, the data may not fully represent the broader Dota2 player base, as player skill and strategies vary across regions and skill tiers. Including data from different regions and skill levels could provide a more complete picture. Furthermore, hero selection was not considered, but could influence outcomes and should be included in future models. The study also excluded player behavior factors, such as communication and decision-making, which likely affect match outcomes. Furthermore, the data only included matches from Captains Mode, whereas Dota 2 offers various other game modes that could exhibit different patterns. Future research could explore these factors for a more holistic understanding of match fairness.

Finally, the beta-binomial model analyzes the overall winning rates for the groups of inexperienced and experienced players. Histograms of these groups' winning rates (Figure 1 & 2) suggest that multimodal models could be used to further investigate individual winning rate distributions within each group.

CONCLUSIONS

Game balance is a critical concern for competitive online platforms, and refining matchmaking systems is an ongoing challenge. Developers must carefully choose metrics that accurately reflect fairness and enhance the player experience. In our analysis, we identified key variables—such as gold, kills, teamfight damage, and experience—that significantly influence match outcomes. The stability and informativeness of our results, as indicated by model diagnostics and posterior distributions, provide confidence in these findings. Future research could expand on this work by incorporating additional variables, exploring nonlinear relationships, and testing interaction effects. The significant variables we identified offer valuable insights that could guide improvements in matchmaking design and serve as a foundation for further exploration.

REFERENCES

How-to: Dota 2 API. (n.d.). http://sharonkuo.me/dota2/matchdetails.html

Dota 2 matches. (2019, November 14). Kaggle. https://www.kaggle.com/datasets/devinanzelmo/dota-2-matches/data