The Curry Effect: Modeling the Ripple of One Player in a 15-Year Dataset

CS109 Project 1: Data Collection.ipynb CS109 Final Project 2: Feature Engineering .ipynb
CS109 Final Project 3: Analysis.ipynb

In 2016, Steph Curry set a record for the NBA's 3-point record for the most in a season with 402. It was also the same year the Warriors set the league record for wins, 73. However, Curry's impact is more than points - it shapes every square inch of the court, changes the behaviors of defenses, and thus ultimately determines whether the Warriors win or lose. I wanted to model that. This project reviews 15 seasons of Golden State Warriors data to ask one question: How much does Steph actually matter?

Data & Framing:

Leveraging play-by-play and game-level NBA data, I filtered for whether Steph Curry played each game and correlated it to game outcomes, opponent strength, and basic stats (an opponent is considered "strong" if their W/L that season is >= 0.6). I treated win probability as the main dependent variable to predict, and split the data by season to see how things have changed over time. The classifier model has 18 features - most of Steph's individual stats, opponent strength, and home or away.

Analysis and Modeling:

The first graph (Fig. 1) shows the Warriors' win percentage when Steph was present based on two conditions: opponent strength and the outcome of a game, as either home or away. To visualize this, we grouped the dataset in three binary ways - Steph, home or away, and opponent strength, and calculated the win percentage for each combination of those conditions. This provided us with a bifurcated view of where Steph's impact is amplified depending on which combination presents itself during the sub-context. For example, when Steph is in the lineup, the Warriors win over 70% of their home games against a strong opponent, but as soon as he is not in the lineup (absent from the study), that win percentage drops profusely, suggesting that his presence above-all-others, is especially pertinent during high-stakes situations. In all three sub-contexts, the Warriors win approximately 20-30% more often when Steph is involved. This clearly helps the case for this project in citing that Steph's value is not just broadly impactful — it builds under pressure.

The second graph (Fig. 2) plots the win rates without Steph and Steph's win rates separately, detailing how the impact of Steph changed over time. This was computed by grouping the dataset by season and separating whether he played or not. This graphic displayed an even better longitudinal look at how Steph's impact has changed, season over season, as his teammates changed. We can once again ascertain that Steph, across most of his career contributes about a sizeable 20-30% additional contribution to the Warriors win rate, but now, we can clearly see how that number changed each season as he acquired superstars like KD, or suffered long term injuries like in 2019 (I did incorrectly mention in the video that he was out for most of the season for an ankle injury, since I was rushing - he was actually out for a broken hand from the 4th game of the season).

The third graph (Fig. 3) simulates a fictional scope of the world in which Steph played every game. I essentially looped through each NBA season, and for games where Steph played, I found the average win rate derived from games grouped by Home vs. Away and Strong vs. Weak opponents. I stored these average win rates in a lookup dictionary based on the (Home, Strong) situations. From there, I simulated 1,000 alternate timelines for each season. For each timeline, all games (including all games Steph missed) were simulated using the average win probability from game conditions that Steph did play. For instance, if Steph missed a home game against a strong opponent, we would use the win probability based on past games Steph did play, which were also home games with a strong opponent. Each simulated game was treated as a Bernoulli trial, where the outcome was either a win with probability p from the lookup table. These simple simulations generated a distribution of possible win percentages per season, of which I also averaged to obtain the simulated win probabilities.

Finally, I trained an 18-variable machine learning model based on a random forest classifier. This model only included Steph Curry's personal statistics in the box score (points, assists, field goals, and free throws, plus-minus, etc.), in combination with home/away and opposition strength. It did not incorporate the statistics of the rest of the team, which was intentional - I wanted to measure Steph's own measurable contribution. This model was trained with a GridSearchCV pipeline to vet the best hyperparameters and preprocessing procedures. The results (Fig. 4) were a surprise - the season-to-season accuracy was surprisingly strong! In most years, the predicted win rates were within 1-3 percentage points of the actual win rate. The slight underpredictions were in seasons when there was increased star talent added (for example, Durant), but overall, this was a sense of comfort that Steph's performance captures a significant portion of the team outcome. This is a mathematical validation of what fans often feel - the gold standard in emotion against Steph's individual performance... if Steph is playing well, the Warriors win.

Conclusion

The evidence paints a strong picture: Steph Curry is not just a good player; he is a key piece of the Warriors' success. The drop in expected win percentage when Curry did not play, the change in expected wins if Curry played every game, and the ability of the model to predict every game relying only on Curry's stats are all supportive of Curry's importance. The analysis by season demonstrates how his value remained consistent while the players around him, whether teammates or opponents, changed. The modeling provides support for a belief espoused by fans: if Curry plays, and plays well, the Warriors are fundamentally a different team. This conclusion is supported by numbers, not highlights and fluff, and numbers and rigorous probability that provide a much more lucid, measurable picture of Curry's generational impact.

Significance

Outside of basketball, this work demonstrates how probability and data modeling are able to measure the unseen ripple effects of a single participant in a complex system. The methods discussed here could pertain to many other domains - for example, the contribution of a teacher in a classroom, a lead engineer in a development team, or a change in policy towards public health. The larger question is: how does one put a value on influence.

Appendix

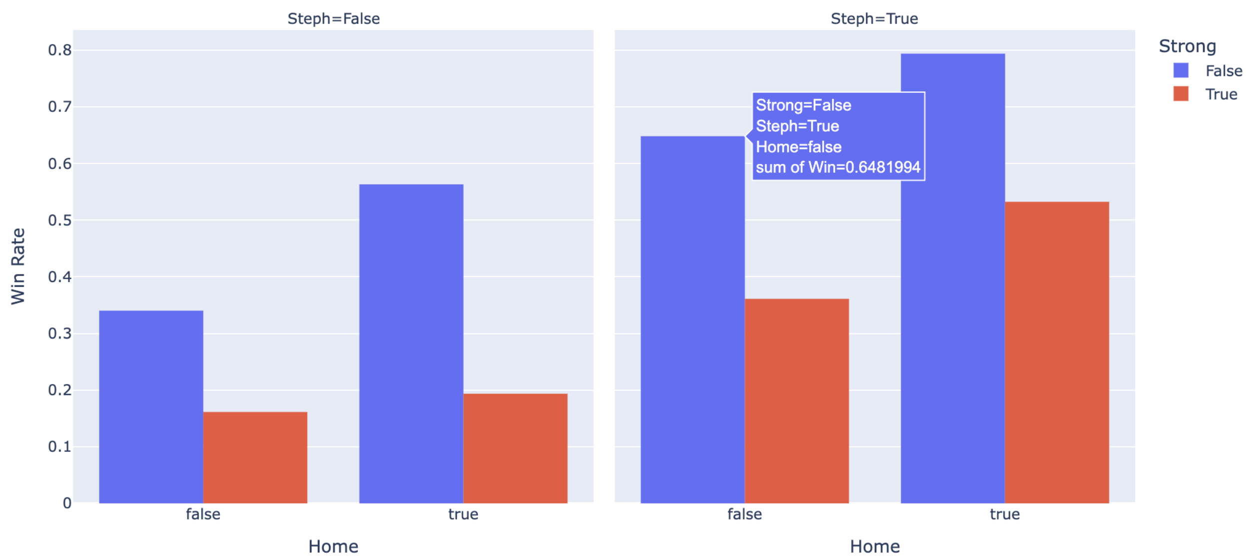

Fig 1. Warriors Win Rate Conditioned on Steph's Presence, Home vs. Away, and Opponents' Strength

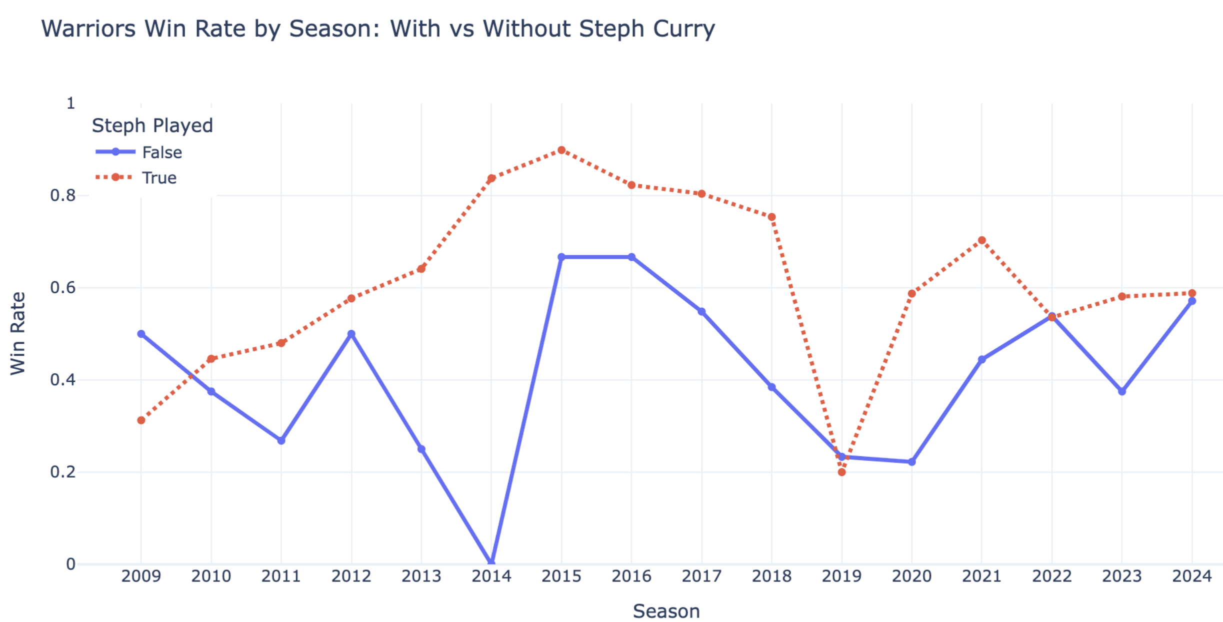

Fig 2. Warriors Win Rate each Season Conditioned on Steph's Presence

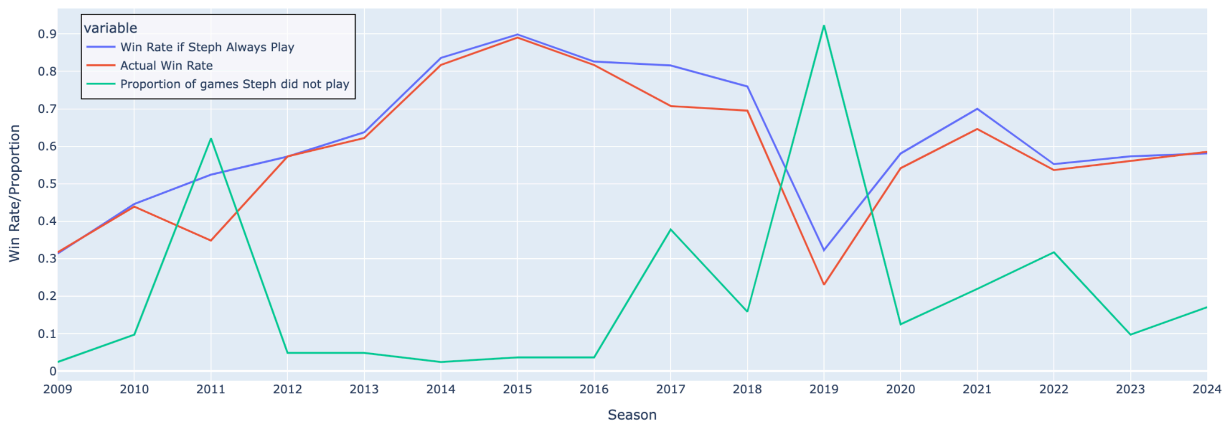

Fig 3. Simulation if Steph Always Play vs. Actual Win Rate and Steph's Absence

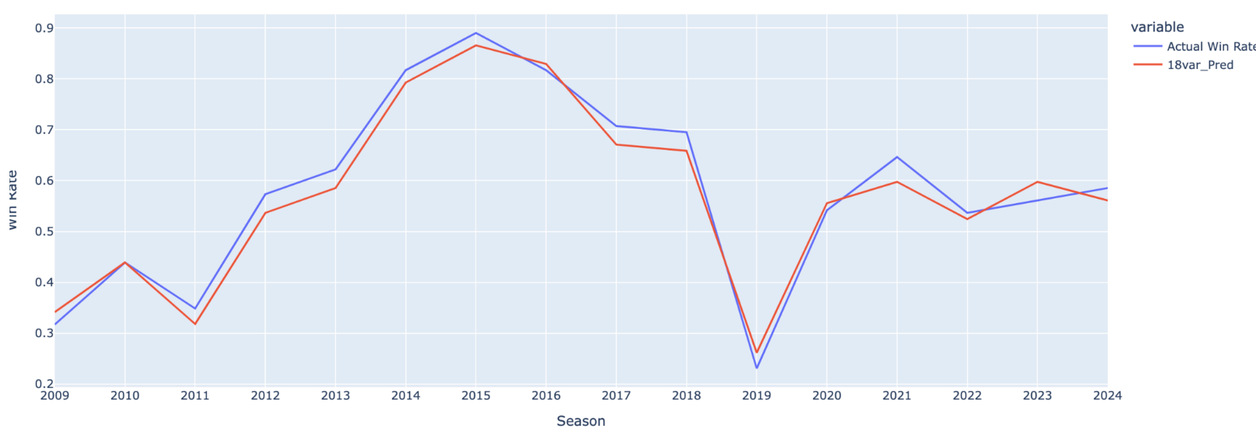

Fig 4. Prediction of Warriors' Performance each Season Solely Based on Steph's Stats and Home vs. Away and Opponent's Strength