Do hot streaks exist in sports?

Hasan Rahman

Abstract

Consider a football team that has won its previous three matches. How does this winning streak affect the likelihood of winning the upcoming match? Is there a measurable impact, and how does it influence the team's chances of victory? In this paper, we aim to examine this relationship and determine whether the results are statistically significant. We will also explore potential explanations for these findings in later sections.

Contents

1 Introduction and overview

1.1 History of the problem

During the Paleolithic period in Mesopotamia, the earliest six-sided dice date back to about 3000 BC. In China, gambling houses were widespread in the first millennium BC, and betting on fighting animals was common. Lotto games and dominoes appeared in China as early as the 10th century.

Figure 1: Bone Dice 5-3rd Century BC in Greece[1]

Later, it became popular to bet on events, which could cover a myriad of different things, such as horse racing, competitive games, or even the outcomes of gladiator games! It seems that people have been attracted to games of chance since even before written history. While seemingly not very practical, many financial instruments are based on gambling. For example, perhaps the oldest form of gambling involved loaning money, with the interest serving as the reward for taking the risk of lending the capital.

During the English Civil War, people even bet on who would win the conflict. Would the Cavaliers or the Roundheads prevail? The modern-day equivalent of this is people betting on platforms like Polymarket for events such as who will be the US president and whether Assad, the leader of Syria, will remain in power by the end of 2024.

Figure 2: Horse race betting in the UK[2]

In England betting would be fairly crude, with most bets being even. In other words if you made a bet on if event X happens, and you spent Y money. If event X did indeed happen you would walk away with $2X$ money. There is an obvious flaw with this system in that it relies on the belief that the probability of event X happening is 50% which is rarely the case.

Therefore, there was a need for bookies to come up with a way to determine the probability of an event occurring so that they could offer appropriate returns. By using past results and data from previous matches, predictive models can be created to calculate the probabilities of various events. These models can go far beyond simply assessing whether team A will win a match; they can also calculate odds for questions such as whether a team will score more than five goals. Using Bayesian statistics, these models can even update odds while the match is in progress. For example, can team A still win despite the fact that it is losing by one point in the first half?

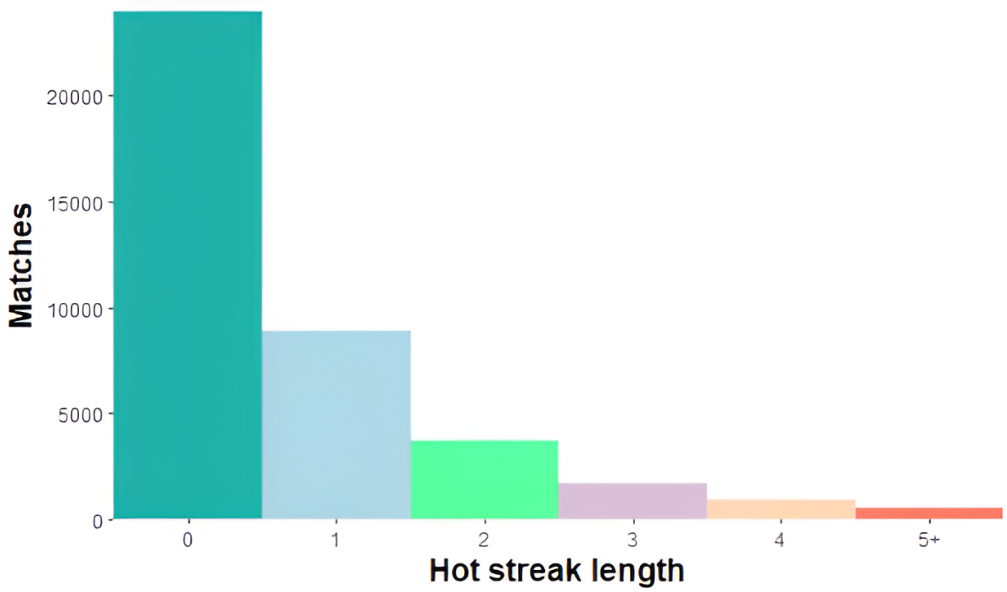
Does a team's winning streak meaningfully impact the probability of winning the next match? Answering this question can help develop better models for predictive events, which will enable more accurate odds and, therefore, more competitive rates. For example, better predictive models for car insurance would allow companies to assess the risk of a customer more easily, enabling them to charge less, as there is less uncertainty associated with the insurance policy.

2 Analysing the data

2.1 Aquiring the data

To prove or disprove whether hot streaks exist in sports, we need to see if the data supports this hypothesis. We can obtain match results from the Premier League from 1993 to 2023 through Kaggle [3]. This database contains thousands of games played by teams across the UK, making it ideal for data analysis due to its ample quantity and good quality, as there is no missing data.

2.2 Proportion of hotstreaks in matches



A hot streak can be defined as consecutive wins in matches. From the graph, we find that out of 24,002 games, there are only 8,918 hot streaks of one match, indicating that only 37.2% of games result in a win for either team. This is due to a large proportion of games ending in a draw. The graph also shows that the longer the hot streak, the more likely it is to continue.

Hot streak table

	Total	Hot streak of 1	Hot streak of 2	Hot streak of 3	Hot streak of 4	Hot streak of 5+
Matches	24002	8918	3657	1712	882	496

This table will count a hot streak that ends after matches, as both a hot streak of 1 and a hot streak of 2.

End of hot streak table:

Hot streak ends	0	1	2	3	4	5+
Matches	15083	5262	1945	830	386	496
Percent of previous	-	34.9%	37.0%	42.7%	46.5%	-

From this table, it becomes clearer that the more games a team has won previously, the more likely they are to win again. This provides evidence for the existence of hot streaks.

3 Creating the model

3.1 Cleaning the data

First, let's create a model to predict whether a team will win or lose the next match. To do this, we need to determine the outcomes of each match for each team. The database provides the home and away goals for each match, which can be extracted using an SQL query on BigQuery.

$$\text{Results for home team} = \begin{cases} 1, & \text{if home goals>away goals.} \\ 0, & \text{otherwise.} \end{cases}$$

$$\text{Results for away team} = \begin{cases} 1, & \text{if away goals>home goals.} \\ 0, & \text{otherwise.} \end{cases}$$

For each team, we can determine whether they won or did not win, which requires different formulas depending on whether they played as the home or away team. We can run another SQL query to find all the matches each team has competed in, whether they were the home or away team, and order the results from oldest to newest.

Additionally, we will create a field that tracks the result of the previous match for each team. For the first match of each team in this database, this value will be null since it is impossible to determine whether they won or lost their last match. In our model, we can exclude these records, as they lack the previous match result and may introduce noise into the model.

3.2 Creating a basic model

To model if a team wins or does not win the next match we can use a generalised linear model. The family will be binomial as match results can only take two values.

$$\begin{aligned} &g(Y_i) = \beta_0 + \beta_1 x_i + \epsilon_i \\ &Y_i = \text{Probability the team wins the next match} \\ &\beta_0 = \text{Intercept} \\ &\beta_1 = \text{Coefficient of predictor} \\ &\epsilon_i = \text{Residual} \\ &x_i = \text{Result from last match} \\ &g(Y_i) = \text{Link function} \end{aligned}$$

3.3 Our results

Putting our model into R we are able to calculate the parameters which lead to the smallest error in our model.

		Estimate	Standard Error	z value	$\Pr(>\|z\|)$
Coefficients table:	β_0(Intercept)	-0.62418	0.01712	-36.450	$< 2e-16$
	β_1(Coefficient of Predictor)	0.26324	0.0275	9.561	$< 2e-16$

Using the link function we can find what the model predicts.

$$P(\text{Win match}|\text{Won last match}) = 41.07\%.$$
$$P(\text{Win match}|\text{Lost last match}) = 34.88\%.$$

3.4 Improving the model

$$g(Y_i) = \beta_0 + \beta_1 x_i + \beta_2 x_{i-1} + \beta_3 x_{i-2} + \beta_4 x_i x_{i-1} + \beta_5 x_i x_{i-1} x_{i-2} + \epsilon_i$$

To improve our predictive model we can add more data, such as did the team win secound last match it competed in.

Coefficients table

	Estimate	Standard Error	z value	$\Pr(>\|z\|)$
β_0(Intercept)	-0.75759	0.02406	-31.490	$< 2e-16$
β_1(Coefficient of Predictor 1)	0.14053	0.03588	3.916	$9e-5$
β_2(Coefficient of Predictor 2)	0.14893	0.03584	4.156	$3.24e-5$
β_3(Coefficient of Predictor 3)	0.22889	0.03057	7.487	$7.07e-14$
β_4(Coefficient of two factor)	0.17774	0.06619	2.685	0.00725
β_5(Coefficient of three factor)	0.12703	0.07340	1.731	0.08351

The two and three factor interactions take into account if the team won the previous 2 or 3 matches. From the results of the table, all betas are significant at the level 10%, with all betas, except β_5 being significant at the 1% level.

$$P(\text{Win match}|\text{Hot streak of 2}) = 42.79\%.$$
$$P(\text{Win match}|\text{Hot streak of 3}) = 51.64\%.$$

4 Using the model

4.1 Conclusion

From the models, we can observe that the relationships are significant and unlikely to have arisen by random chance. Furthermore, there is a notable two-factor interaction related to the outcomes of the previous two matches, significant at the 1% level. This provides strong evidence for the existence of "hot streaks" in sports.

As for the explanations behind these results, they are likely psychological in nature. Players who have just won a match tend to feel more confident and may thus perform better, increasing their chances of winning the next match. Conversely, the opposite is likely true for players who lost their last match, as decreased confidence may negatively affect their performance.

5 References

References

[1] Traveltoeat.(n.d.). *Ancient Board Games. British Museum.* Retrieved from https://traveltoeat.com/ancient-board-games-british-museum/

[2] Onlinebetting.(n.d.). *The History Of Betting On Football* Retrieved from https://www.onlinebetting.org.uk/betting-guides/football/history-of-betting-on-football.html

[3] Kaggle.(n.d.). *Premier League Matches 1993-2023* Retrieved from https://www.kaggle.com/datasets/evangower/premier-league-matches-19922022?resource=download