Do hot streaks exist in sports?

Hasan Rahman

Abstract

Consider a football team that has won the previous 3 matches; how does it affect the likelihood of winning the upcoming match? If so, what is the impact and how does it affect the team's chances of winning? We shall examine this relationship in this paper and attempt to demonstrate whether the results are significant. We'll look at possible explanations for these findings later.

Contents

1 Introduction and overview

1.1 History of the problem

During the Paleolithic period in Mesopotamia the earliest six-sided dice date 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: Knuckle Bone Dice 5-3rd Century BC in Greece[1]

Later, it would become popular to bet on events. Which could cover a large myriad of different things such as horse racing, competitive games or even the outcome 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 gambling was loaning money with the interest being the reward for taking the risk of lending the capital.

During the English Civil war people even betted on who would win the war. Would the cavaliers or roundheads win? The modern-day equivalent of this is people betting on polymarket for events like who will be the US president and will Assad leader of Syria stay 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 an event occurs, so they can give appropriate returns. Using past results, and data from those previous matches predictive models can be made to calculate probabilities of events. These models can go far beyond will team A win a match, it can calculate odds for will team score more then 5 goals? Using Bayesian statistics it can calculate odds while the match is playing. Will team A win despite the fact that it is losing in by 1 point in the first half?

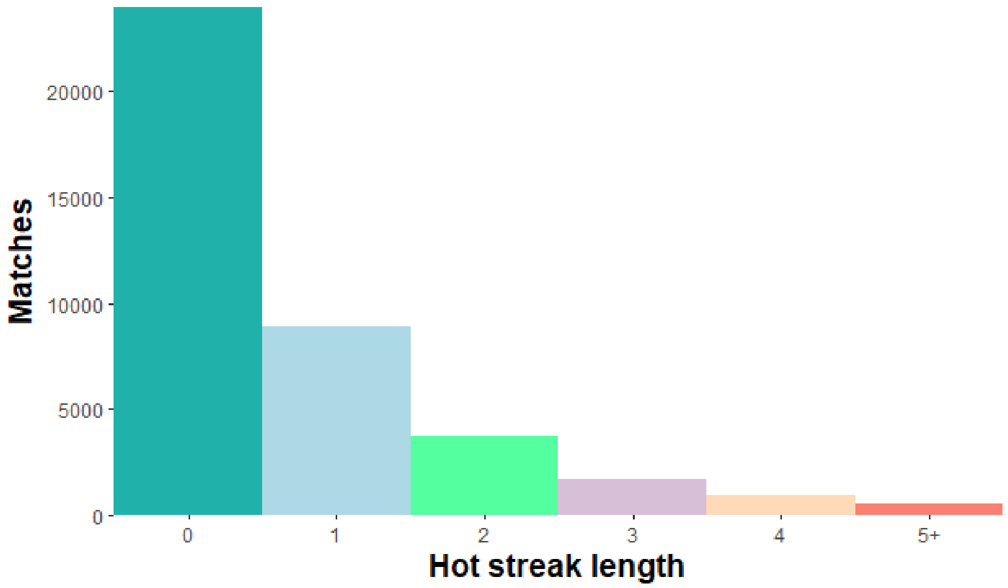
Does a teams winning streak meaningfully impact the probability of winning the next match answering this question can help develop better models for predictive events. Which will allow models to give more accurate odds and therefore more competitive rates. For example better predictive models for car insurance would mean companies would be able to access the risk of a customer more easily. Allowing them to charge them less, as there is less uncertainty with the insurance policy.

2 Analysing the data

2.1 Aquiring the data

To prove or disprove if hot streaks exist in sports, we will need to see if the data backs this hypothesis. We can get from Kaggle the results of matches in the Premier. League from 1993 to 2023[3]. This database has thousands of games, played from teams over all of the Uk. Making it ideal for data analysis as there is enough data and it is of 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 the 24002 games, there are only 8918 hotstreaks of 1, showing only 37.2% games result in either team winning, this is because a large proportion of games result in a draw. From the graph it shows that the longer the hotstreak is the more likely it will 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 more clear, that the more games a team has won before the more likely they are to win another. Showing evidence for the existence of hot streaks.

3 Creating the model

3.1 Cleaning the data

First, let's create a model to predict if a team will or won't win the next match. To do this we will need to determine when each team won or lost a match. The database provides the home and away goals for each match. This can be added through the use of an SQL query, on big query.

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

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

For each team it can be determined if they won or didn't win. Which requires a different formula if they are playing as the home or away team. Another SQL query can then be run, to find all the matches each team competed in, either when they were the home team or the away team. It will then order from the oldest to newest. We create another field that tracks the previous match result for that team. The first match for determine each team has in this database this value will be null, as it is impossible to determine if they did or did not win the last match. In our model we can exclude these records as they do not have the result for the previous match and would add noise to 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.

Coefficients table:

	Estimate	Standard Error	z value	$\Pr(>\lvert z\rvert)$
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.

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

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 see the relationship is significant, and unlikely to be generated from random chance. Additionally, there is a significant two-factor interaction from the previous two matches in 1% this shows strong evidence for hot streaks existing in sports. As for explanations for the result it is likely physiological, players who have just won a match are more confident and will therefore perform better increasing the probability of winning the next match, while the reverse is true if you lost the last match.

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