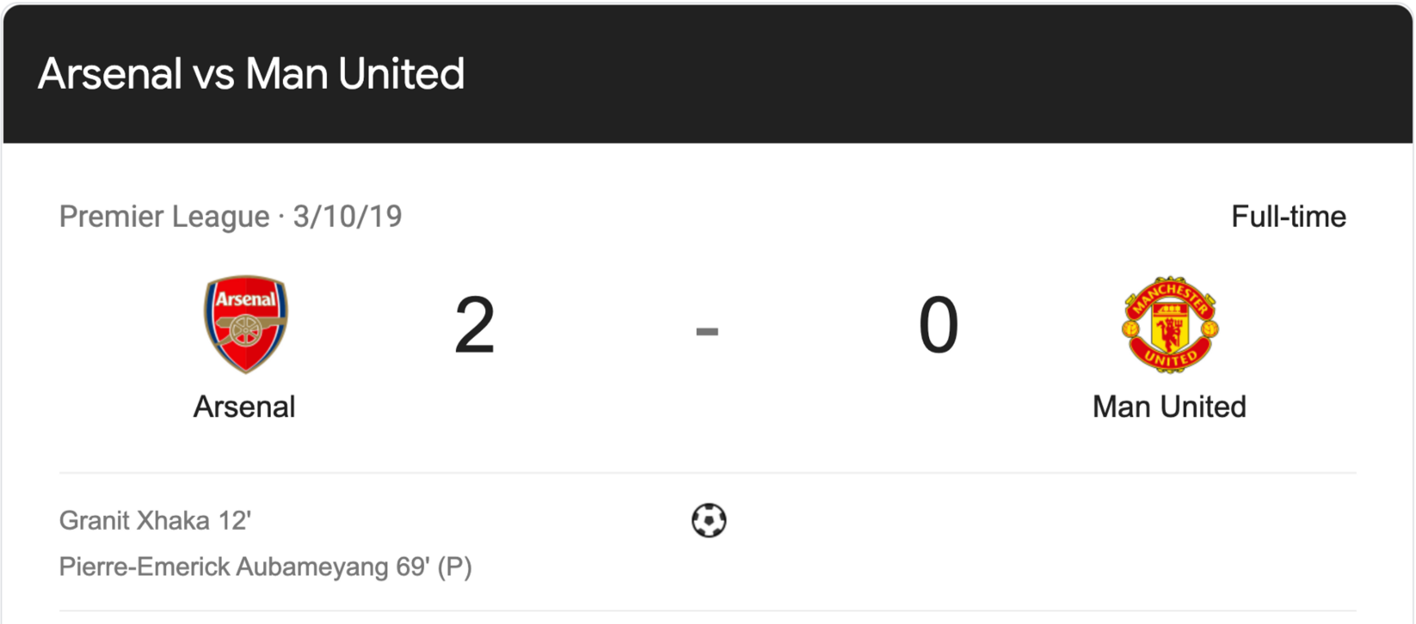
Towards Data Science

‘Making big bucks’ with a data-driven sports betting strategy

I was watching the match between Arsenal and Manchester United last weekend, one in which the home side was generally regarded as an underdog.



Arsenal vs Man United

Premier League · 3/10/19

Full-time

Arsenal 2 - 0 Man United

Granit Xhaka 12'

Pierre-Emerick Aubameyang 69' (P)

To everyone’s surprise, Arsenal came out on top. It really could have gone either way. United hit the woodwork twice in the first half. But a rare David De Gea misjudgment from Xhaka’s swerving shot and a generously awarded penalty added another unpredictable result in this confounding topsy-turvy season.

And did I mention that Tottenham Hotspur was beaten by Southampton the same weekend?

As another round of surprising results from the Premier League unfolded, I kept thinking about the algorithm I developed. Would it be able to correctly predict the results on a consistent basis? There is some inherent randomness in the model, but is it enough to factor for the tantalizing poised nature of the PL, where relegation-zoned Southampton clinched a victory against all-star Tottenham?

So I decided to bring it back and back-test.

One of the difficulties of testing an algorithm is to find a good benchmark for its performance. Say, if my prediction has an accuracy of 50% over 200 matches, is it good, bad or mediocre? It surely outperforms random guessing (with equal probability of 1/3 for Win, Draw and Lose), but it does not sound that great, does it?

How about comparing my results to professional football pundits?

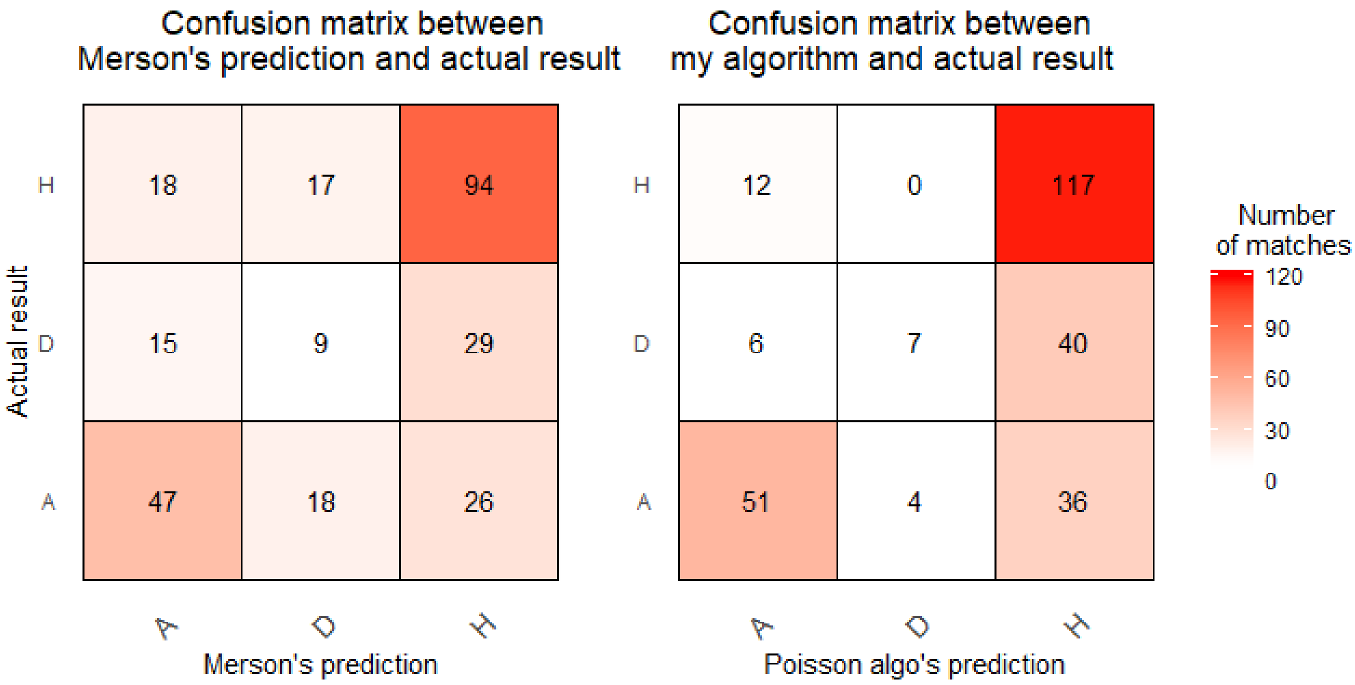
So I found out that every week, SkySports website published a prediction for that week fixtures by Paul Merson [1], an ex-Arsenal-player-turned-pundit who had won several titles.

I’m honestly not a big fan of Paul Merson, after what I thought was his relentless criticism against his former club.

Just listen to what Arsenal former manager, Wenger had to say about him:

These debates that I hear are a joke, a farce. People [Merson] who have managed zero games, they teach everybody how you should behave. It’s a farce.

Nonetheless, this is a gold mine for me, because I can now compare my algorithm against an ‘expert’. No matter what your opinion about him, the prediction of an ex-Arsenal player for the Arsenal-Man United match will surely be more dependable than an obscure model that runs on randomly spitting out numbers.



Here, I compared the results between 273 matches Merson predicted this season. He achieved a **54.9% accuracy**, while my Poisson-process algorithm achieved a surprising **64.1% accuracy**.

Interestingly, Merson predicted a 2–2 draw between Arsenal and Manchester United, saying " both teams will have a go at each other and there will be goals." My algorithms, by averaging the number of goals Arsenal scored and conceded at home, assigning a slight edge and winning probability of 45% to Arsenal, comparing to 27% to Man United.

The result startled me. A 10% edge over an expert's opinion is huge. And I did not even have to do much besides asking the beloved Poisson processes to chunk out numbers.

This is when I started looking into sports betting. And I enter a new game against a new opponent: it's me against the bookies.

If you ever think that the terms and quoted APR on your credit cards are complicated, try venturing into those betting websites once. They are just plain crazy.

Take the US Odds for example. If you see an odds of+300, it means your payoff is $300 if you bet 100 and win. This is fine, but then they have **negative odds**, like an-150 odds. What the @#*!$% is that? It means in order to make a $100 profit, you'll need to place a $150 bet. So, US odds are a number greater than or equal to 100, sometimes preceded by a + to indicate the number is your profit, sometimes preceded by a — to indicate the amount you need to bet to win $100.

I mean, they are still using Feet and Fahrenheit anyway

GLOBAL CONFLICT BETTING ODDS
ODDS
WAR WITH NUCLEAR ARMED RUSSIA 7 TO 4
WAR WITH NUCLEAR ARMED CHINA 5 TO 2
THAT COCKROACHES WILL INHERIT THE PLANET 2 TO 1
INKCINCT

2016-564 © INKCINCT Cartoons www.inkcinct.com.au

For the purpose of this project, we will use a nicer system: **the European Odds**. It's simple: they tell me how much I will get back if I bet $1. For example, Bet365 gives an odds of 2.4 for the event that Arsenal beating Man United, 3.6 for a draw and 3 for Manu winning. This means that I would have come out of the bet with $2.4 (a $1.4 profit) in my pocket if I had put a $1 bet for Arsenal.

But things are not always nice and simple. In reality, to maximize profit, bookmakers employ teams of data scientists to analyze decades of sports data and develop highly accurate models for predicting the outcome of sports events and giving odds to their advantage.

Let's assume that the bookmakers' odds are a perfect reflection of the probability of the various teams winning, drawing or losing. So, for that Arsenal-Man United clash, since the odds Bet365 gave to Arsenal winning are 2.4, the probability of them winning is simply 1/2.4 = 41.6%, surprisingly close to my prediction of 45%. Similarly, the probability of Man United winning is 1/3.0 = 33.3%, and the probability of a draw is 1/3.6 = 27.8%.

Hang on a minute !!!

41.6% + 33.3% + 27.8% = 102.7%! That's odd (No pun intended!!!)

The reason the probabilities don't add up to 100% is that **the odds aren't fair**. That extra 2.7% is **the bookmaker's advantage**. To get the real probabilities, we need to correct for the profit by dividing through by 102.7. So the bookmakers' true probability of an Arsenal win is 41.6/102.7 = 40.5%, the probability of a United win is 33.3/102.7 = 32.5%, and for a draw, it is 27.8/102.7 = 27.06%. For a perfectly efficient bookmaker, these are the probabilities of each outcome.

Now, this is the funny business: if the odds perfectly reflect reality, then it doesn't matter which outcome I bet on — my expected profit is always the same.

If I bet $1 on Arsenal, I expect to get back :

$$E(\mathrm{W} = \text{Arsenal wins}) = \; p_W \times \text{pay-off} + (1 - p_W) \times 0 = \quad 0.325 \; \times 3 \approx \$0.97$$

The expected profit is the same if I had betted for Man United:

$$E(\mathrm{W} = \text{United wins}) = \; p_W \times \text{pay-off} + (1 - p_W) \times 0 = \quad 0.325 \; \times 3 \approx \$0.97$$

And — you guessed it — if I bet on a draw, I expect to get back 97 cents. **On average, the bookmaker will take about 3 cents from me per $1 bet.**

This understanding does not stop me from trying to exploit any potential inefficiencies in the market. At first, I devise the general bet strategies.

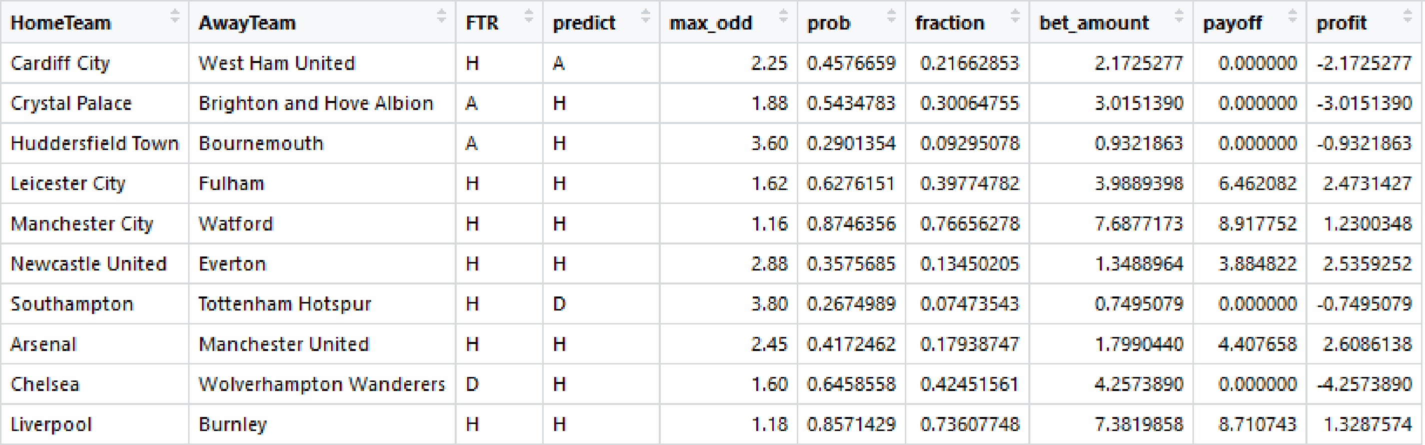
- I set out a budget of $1000, divided equally to 30 previous rounds of the Premier League. So each weekend I have roughly $33 dollars to bet.
- For each match, a prediction will be made by one of the three methods: (a) Paul Merson's prediction, (b) my Poisson process algorithms and (c) a random assignment of equal probability to win, draw and lose.
- With the prediction, I find the **highest odds among 6 online betting houses**. This means if I win, I get the highest profit possible. This will be the odds at which I place my bet.
- For each match, the amount of bet will be calculated by the **Kelly criterion** [2], which works based on the principle: you should invest only a fraction of your wealth. By keeping some aside, you will not end up in bankruptcy. The optimal fraction (f) depends on each individual bet:

$$f = \frac{\text{Edges}}{\text{Odds}} = \frac{p_W^* \times x - (1 - p_W^*)}{x}$$

Implementing the Kelly Criterion is quite simple in R:

The question remains what is considered the true probability of events (p*) in the Kelly criterion's formula. As we have seen in the previous parts, we can take the inverse of the odds given by any specific betting house, but this will not end up great as they are tilted in the house's advantage. However, if we aggregate all the odds from many different betting houses, we should get a better reflection of how bookmakers view the probability of an event, Arsenal defeating Man United for example:

$$p_W^* = \frac{1}{\frac{1}{n} \times \sum_1^n x_i}$$



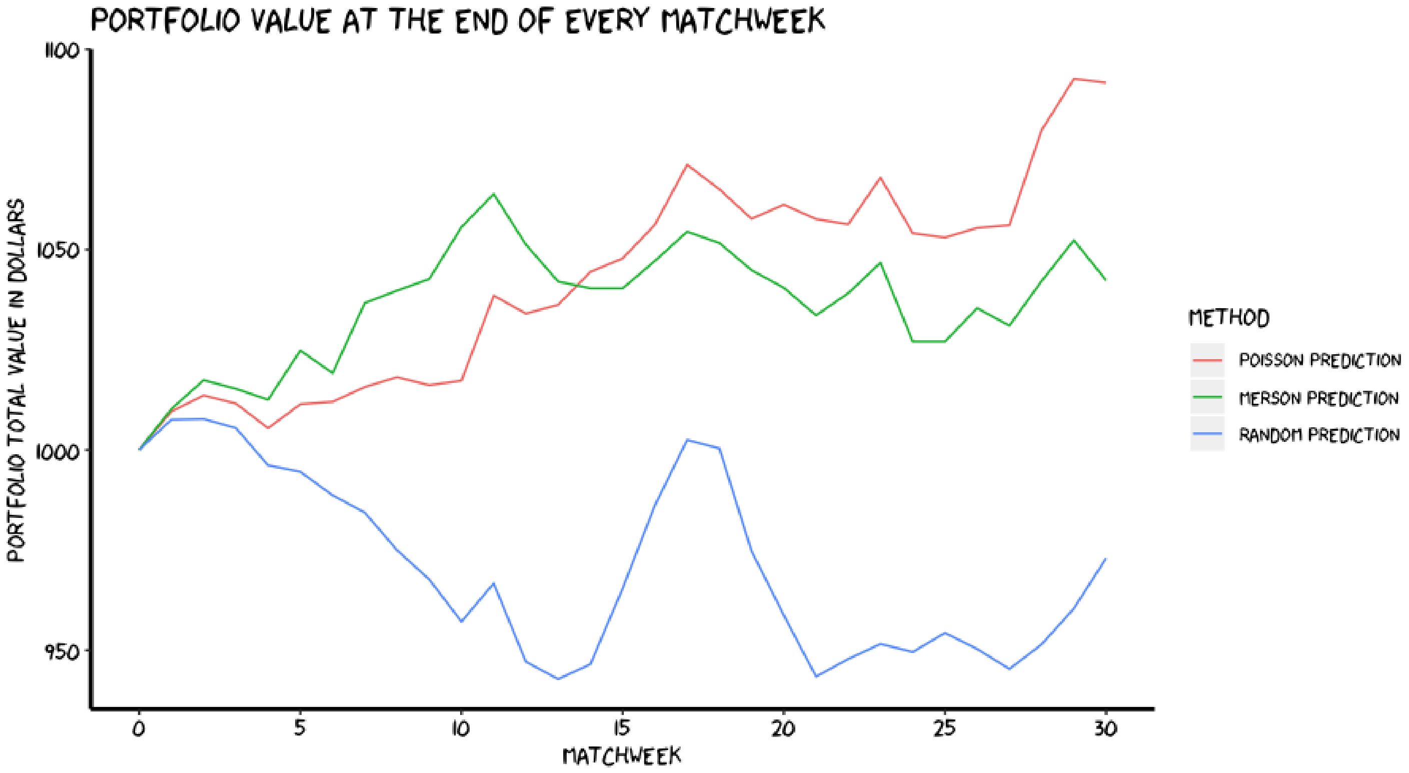
HomeTeam	AwayTeam	FTR	predict	max_odd	prob	fraction	bet_amount	payoff	profit
Cardiff City	West Ham United	H	A	2.25	0.4576659	0.21662853	2.1725277	0.000000	-2.1725277
Crystal Palace	Brighton and Hove Albion	A	H	1.88	0.5434783	0.30064755	3.0151390	0.000000	-3.0151390
Huddersfield Town	Bournemouth	A	H	3.60	0.2901354	0.09295078	0.9321863	0.000000	-0.9321863
Leicester City	Fulham	H	H	1.62	0.6276151	0.39774782	3.9889398	6.462082	2.4731427
Manchester City	Watford	H	H	1.16	0.8746356	0.76656278	7.6877173	8.917752	1.2300348
Newcastle United	Everton	H	H	2.88	0.3575685	0.13450205	1.3488964	3.884822	2.5359252
Southampton	Tottenham Hotspur	H	D	3.80	0.2674989	0.07473543	0.7495079	0.000000	-0.7495079
Arsenal	Manchester United	H	H	2.45	0.4172462	0.17938747	1.7990440	4.407658	2.6086138
Chelsea	Wolverhampton Wanderers	D	H	1.60	0.6458558	0.42451561	4.2573890	0.000000	-4.2573890
Liverpool	Burnley	H	H	1.18	0.8571429	0.73607748	7.3819858	8.710743	1.3287574

For Matchweek 30, with 5 matches predicted correctly and the best odds chosen from 6 houses, we totaled **a net loss of $0.9 or 90c** for this round with the Poisson prediction embedded in our betting strategy. Our biggest loss came from Chelsea's failure to snatch 3 points at home against Wolves.

Now, assuming that I have used this strategy from the very beginning of the Premier League, let's see how quickly we managed to get rich.

Both my algorithm and Merson's predictions -when coupled with the max odd strategies with Kelly criterion, net positive return by the end of Matchweek 30, with the Poisson-process prediction achieving a whopping **9.1%** return with a normalized return of **0.3%** per Matchweek. To put in perspective, the market price return of the Vanguard S&P 500 ETF is **4.6%** [4].

The random method **nets a loss of 19%** on the first iteration, mainly because a few lucky bets here and there (Man United lost to West Ham) cannot compensate for a lot of bad bets (Leicester, Huddersfield won at Etihad, Tottenham lost to Bournemouth, like honestly?). Even if I rerun the random prediction many times, suffice to say that I have seen less than 10% of the cases where the random methods have positive returns.



Obviously, there are inherent risks in this optimal Poisson model. Take Matchweek 24, where we were struck with a net loss of **$14** dollars. Both Merson and the Poisson-process model (and me !!!) was very confident in Liverpool, Man City, Man United, and Chelsea earning 3 points against Leicester, Newcastle, Burnley, Bournemouth respectively, proposing a total bet of **$19.** Result: Liverpool and Man United failed to grab all 3 points while Chelsea and Man City was defeated. All in the same weekend !!!

Before you clone my Github repo and raise capital for your sports hedge fund, I should make it clear that there are no guarantees. You need a large starting capital (I simulate with $1000 but every week I have only $33 to bet), a lot of patience and a cool head.

If anything, this article is a toy example of what you could potentially do. But the bookmakers have made it extremely difficult for anyone to gain sustainable profits. If the bookie thinks the probability of a win is 1/6, then he will guarantee that his expected intake minus payout is positive by setting the odds to be less than 5, maybe something like 4.6. If there are still a lot of people placing a bet at 4.6 odds, then the bookie surely realizes that the probability of a win must be higher than his own estimation and will adjust the odds to say 4. Chances are that by the time the code infers the most optimal odds, it has been changed.

Furthermore, if you do start to make a regular profit, bookmakers can simply thank you for your business, pay out your winnings and cancel your account. This is what has happened to a research group from the University of Tokyo [3].

A few months after we began to place bets with actual money bookmakers started to severely limit our accounts. We had some of our bets limited in the stake amount we could lay and bookmakers sometimes required “manual inspection” of our wagers before accepting them