Fool’s Gold in Football Transfers

Making reliable sense from unreliable data

Motivation:

In the context of the football transfer market, we want to develop heuristics and metrics to distinguish, as much as we are capable, reliable information and reliable sources, from unreliable information and unreliable sources

As it relates to business value, it is useful to emphasize that our motivation is almost entirely about the information angle. There is little opportunity to leverage information about the transfer market for business purposes

However, in general terms, the capacity to accurately determine credibility of information is a monumentally valuable asset, with possible applications in finance, recruiting, defense and intelligence, corporate strategy and politics, just to name a few

Background:

Roster composition for football teams in the EPL (English Premier League) and other major leagues around the world is largely determined by the transfer market. In the transfer market, a player from one club is sold to another for an amount of money agreed on between the two clubs

Accordingly, there is a huge amount of interest from football fans in the transfer market, especially during the offseason. From that, there is a large number of reports, speculation, rumours, and other information from the football press and other purveyors of information to address this demand

Data:

Chelsea Football Club, summer window 2020

Major acquisitions:

Kai Havertz
Timo Werner
Ben Chilwell
Hakim Ziyech
Edouard Mendy
Thiago Silva

Targets not acquired:

Sergio Reguilon
Jadon Sancho
Moussa Dembele
Jan Oblak
Kalidou Koulibaly

Raphael Varane
Nicolas Tagliafico
Declan Rice
Dean Henderson

Media

The Guardian (UK) football section, Apr 1, 2020 to Oct 31, 2020

The Guardian 2020 Top 100 Male Footballers (all ballots)

https://docs.google.com/spreadsheets/d/1_UfNkj95y3wgfdsxsEC4tcdlRfTts8c5mVLHVIfdSOs/edit#gid=3

Methodology:

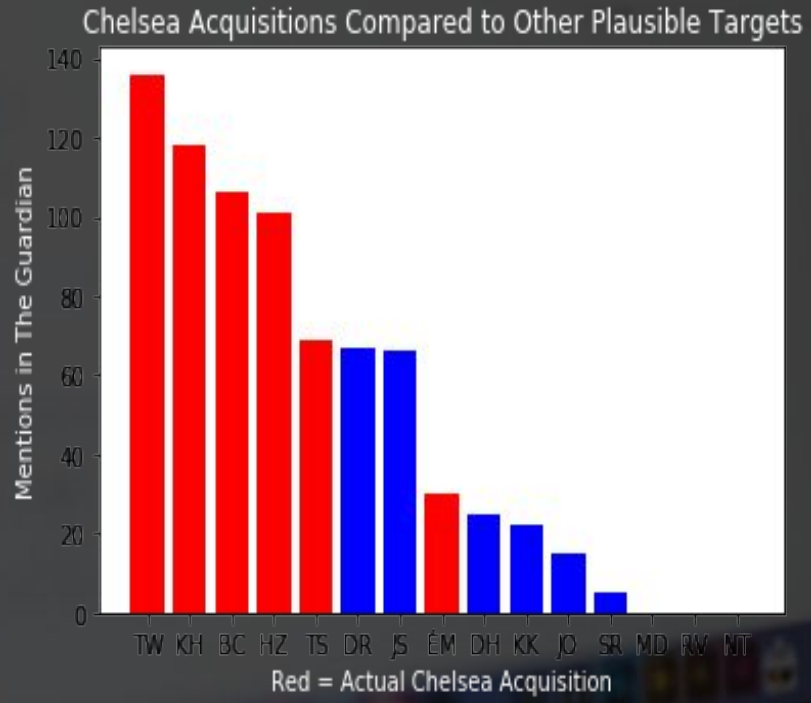
We query every player that Chelsea bought in summer 2020 against every piece of content published in The Guardian Football section. As a first approximation, we consider it to be a match if the string “Chelsea” and the player’s name are both in the same piece of content

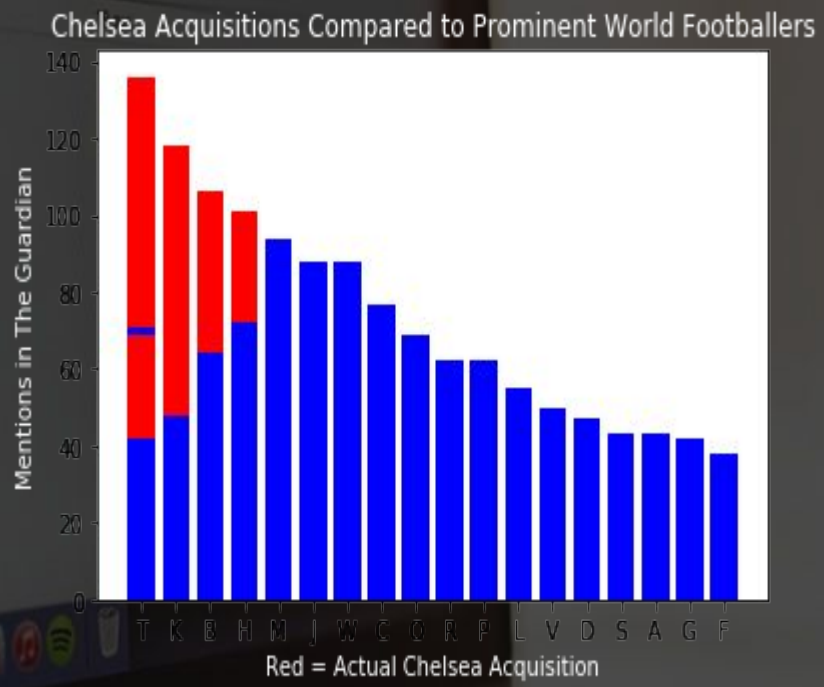
Then, we compare those matches against a similar methodology for a small number of other players that Chelsea was known to be considering. Seeing promising results, we extended these queries to a larger list of 440 of the world’s best footballers

The guiding logic is the hope that if The Guardian repeatedly mentions a given player significantly more often than an otherwise comparable player, it is a positive indication that Chelsea will acquire that player.

Results:

And in fact, that's exactly what we found.

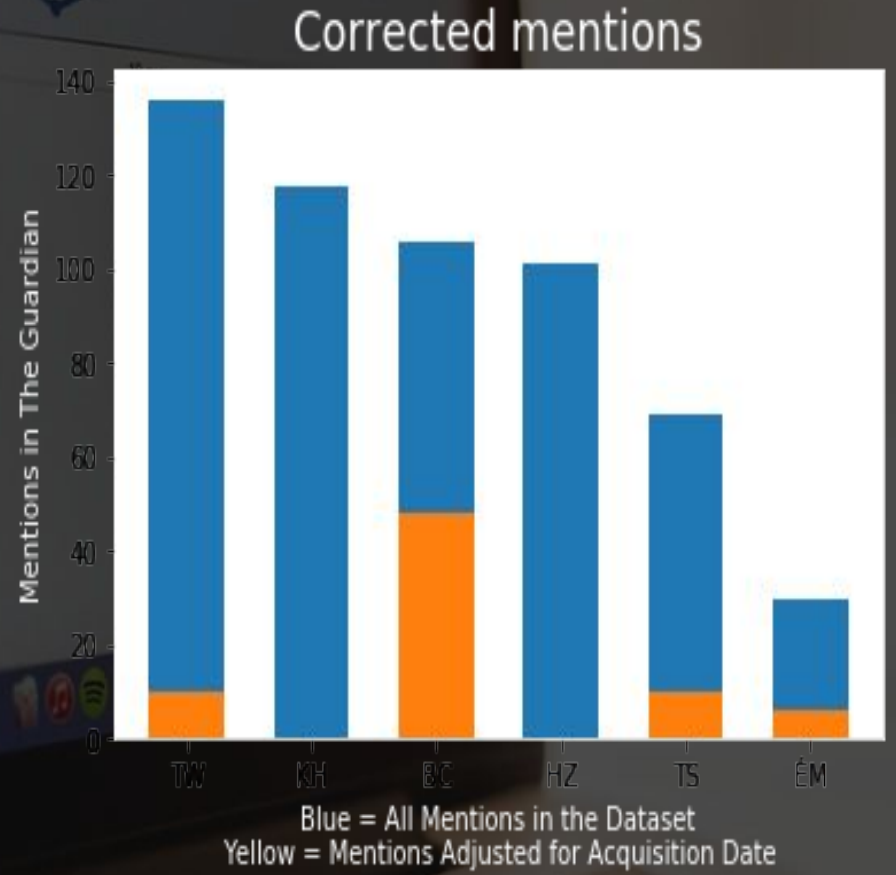




Unfortunately:

This seemingly happy picture hides a severe methodological error. We counted as a hit, all mentions of the acquired players in the dataset. However, each of these six players actually were acquired by Chelsea. And most of the mentions could and were from content published after the acquisition.

If we correct the mentions to those in the dataset made 4 days before the acquisition or prior, many and in couple of cases all of them disappear, thus negating our prior encouraging results.



Things To Do Differently:

The easiest improvement is to change the date range of the data. Given the timing of Chelsea's acquisitions, it was important to get earlier data. And, the October data at the end was probably not very meaningful at best and probably should be omitted.

More fundamentally, there must be NLP or some much more sophisticated textual analysis on the data to generate reliable information. That was in fact my intention for the project, except that it (erroneously) seemed as though I found what I wanted without it.

Areas to Extend and Advance:

To improve the project, it would be valuable to get data from other data sources, specifically lots of sources, other establishment press and social media, for a narrow time period where transfer activity is likely to be concentrated.

I'd also want to speak with a football executive or a transfer-savvy Chelsea supporter. Because there are not that many acquisitions for any given summer, we are not aiming for statistical significance. Instead, it is a more plausible intention to create insight that organically expands the existing knowledge base for experts.