Massive Data: A Goalscoring Model More Predictive Than Bookmakers Odds Implications

Both Bookmaker Odds and recent xG data are the trusted go to indicators of goalscoring potential for many FPL managers. In my previous analysis on how effectively xG data can reflect true player level goalscoring potential I promised to share an comparison of the predictive power of these data sources. This is exactly what this post will assess, but there is a twist- I will also introduce implications from Spread Markets (used in the Implied Odds planner) and FPL Reviews own Massive Data model.

In the previous post I had described an ideal model- which effectively describes the basis of what became the Massive Data model:

> *A smart model might look at all past seasons, giving higher weightings to fresher & greater sample data while giving reference to the team quality & the players role (position and set-piece hierarchy) in the team at the time- and then consider the players current situation. It might also refresh after each and every match to maintain a live rating- though that is really for people who want to spend significant time on this.*

This ended up taking several months of development and testing which in turn delayed this analysis quite a bit as I wanted to create something to really stress test the value of implied odds data.

The Models Under Analysis

In this post I'm going to run through an analysis of a few different models to determine which is most effective and create a fair benchmark so we can say which systems are clearly good or bad. It's worth considering that in isolation no normal person would have an idea what a good R^2, RMSE or AUC score would be for this problem. Below are the models included in the analysis:

- Implications from Bookmaker Scorer Odds (ie. margins removed)
- Implications from Spread Markets Scorer Odds
  FPL Review Massive Data Model
- xG90 data from equivalent of last 5 full matches

- Goals per 90 minutes (G90) data from equivalent of last 5 full matches
- Basic FPL Position/Fixture Difficulty based Model

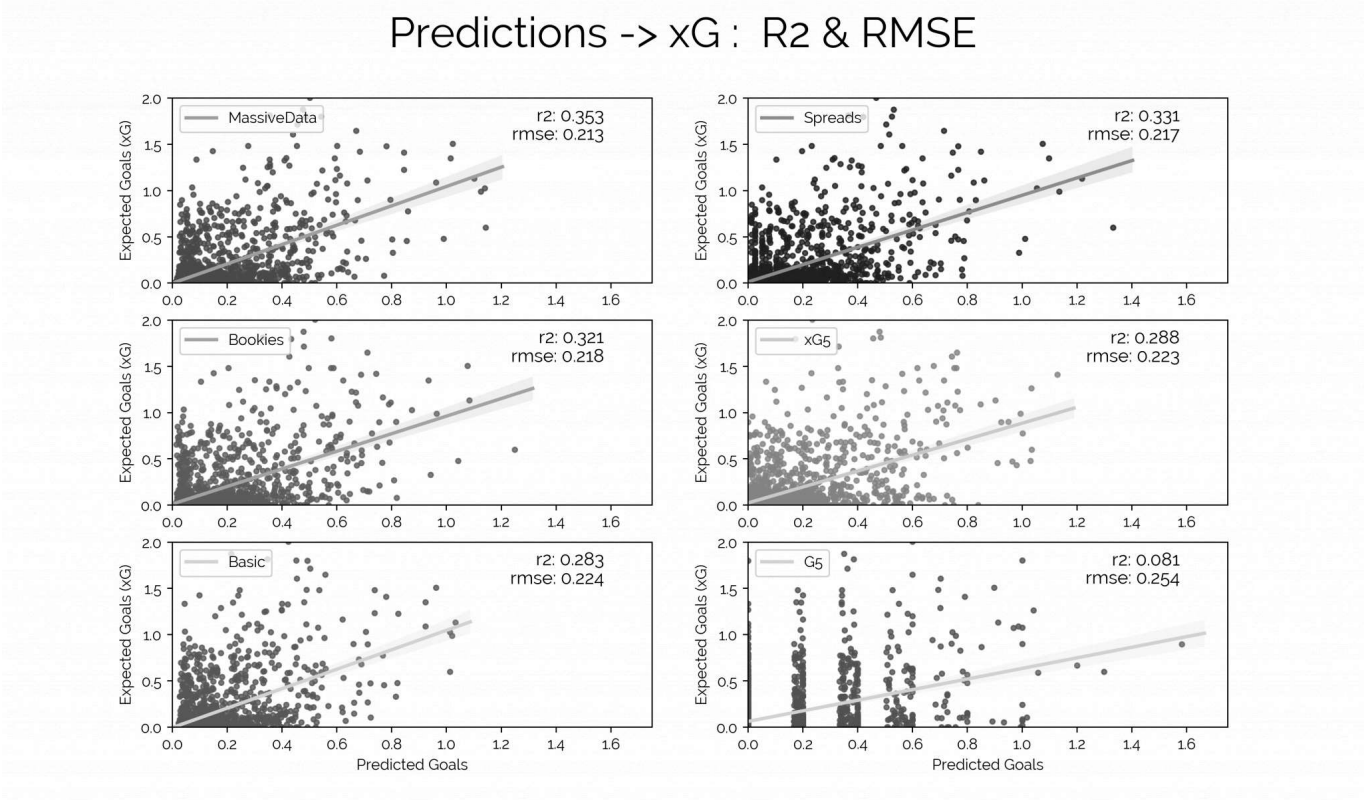
2019/2020 Model Prediction Performance

Below is the description of the sample used- it's vital the same sample is applied across all prediction models for a meaningful comparison.

- Excludes GKs (inflate prediction scores- anyone can predict a GK won't score)
- Data is for GW9-26, as did not have Spread Odds prior to this
- Minimum 450 minutes background data (for xG90 model)
- Only samples who played 90 minutes are included (avoids distortion from subs & keeps things neat)
- Only players I had Spread Odds and Bookmaker Odds for are included

This left me with 2,021 samples each with unique predictions from all models, post-match quality adjusted xG and goal data. Using inconsistent samples would render this analysis almost worthless.

So lets dive right into it and see which of the models predicts quality adjusted xG most effectively. The two metrics I am using to analyse predictive performance are R^2 and RMSE (root mean square error)- these are typical metrics for judging statistical models and probably the two most popular for this kind of data. R^2 indicates how well the predictions fit with the results, a high score is desirable. RMSE is a measure of prediction error, a low score is desirable.

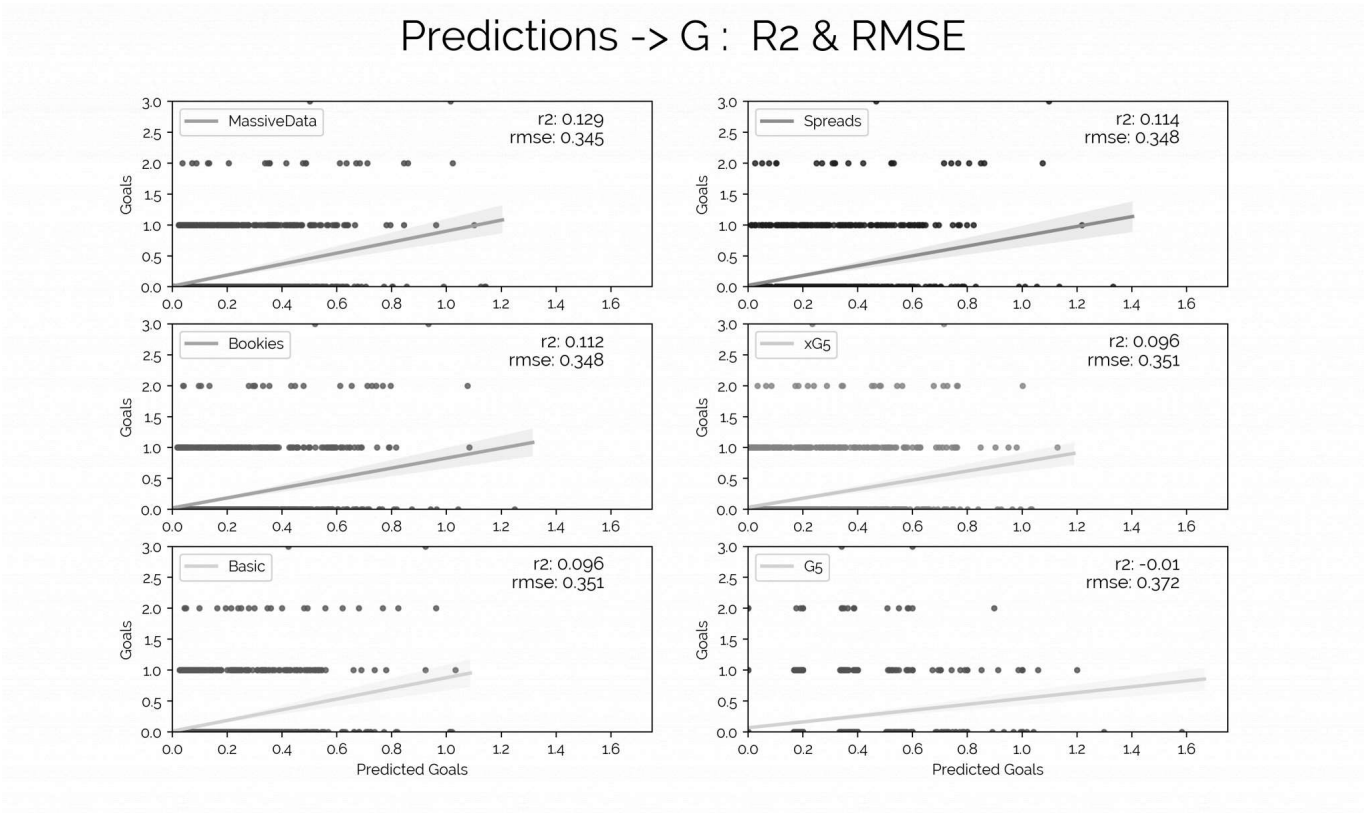


Scatter plots for each model against the resulting xG

Based on the agreement between the R^2 and RMSE results above the following order would is the ranking of the models effectiveness:

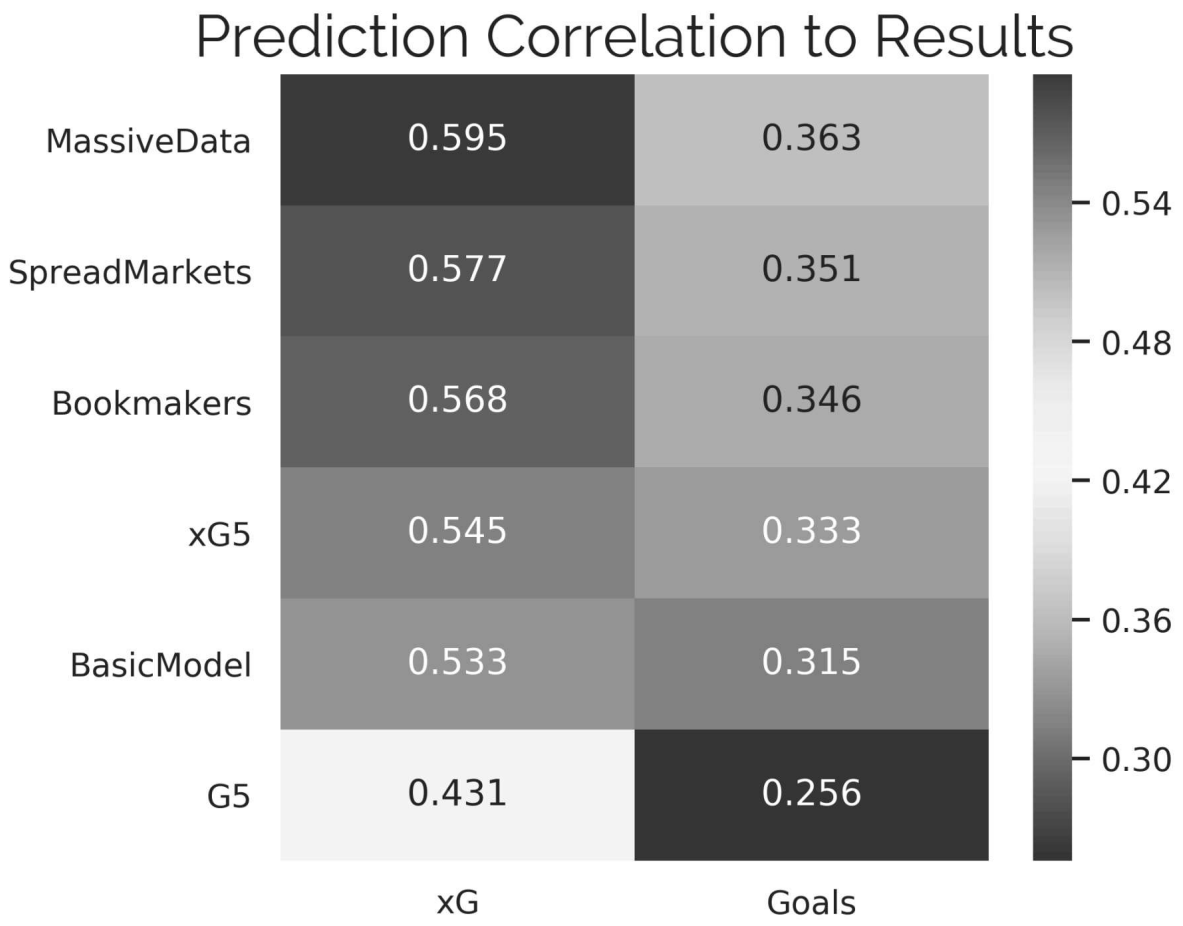
Rank	Model	xG: R^2	xG: RMSE	G: R^2	G: RMSE
1	Massive Data	0.353	0.213	0.129	0.345
2	Spread Markets	0.331	0.217	0.114	0.348
3	Bookmakers	0.321	0.218	0.112	0.348
4	xG5	0.288	0.223	0.098	0.351
5	Basic Model	0.283	0.224	0.096	0.351
6	G5	0.081	0.254	-0.01	0.372

My preference is to model against quality adjusted xG data however some others will rather checking against goals scored and it's worth doing as a confirmation of the initial results. The results in terms of goals retain the same indication of predictive power as was found in terms of predicting xG.



Scatter plots for each model against the resulting Goals scored

A simpler way to view the results can be seen below- this is the correlation of each prediction model with the resulting goals and xG- the results are in line with the above plots. We can also see how xG is more predictable while there is an added level of variance in converting chances to goals, which goals harder to predict.

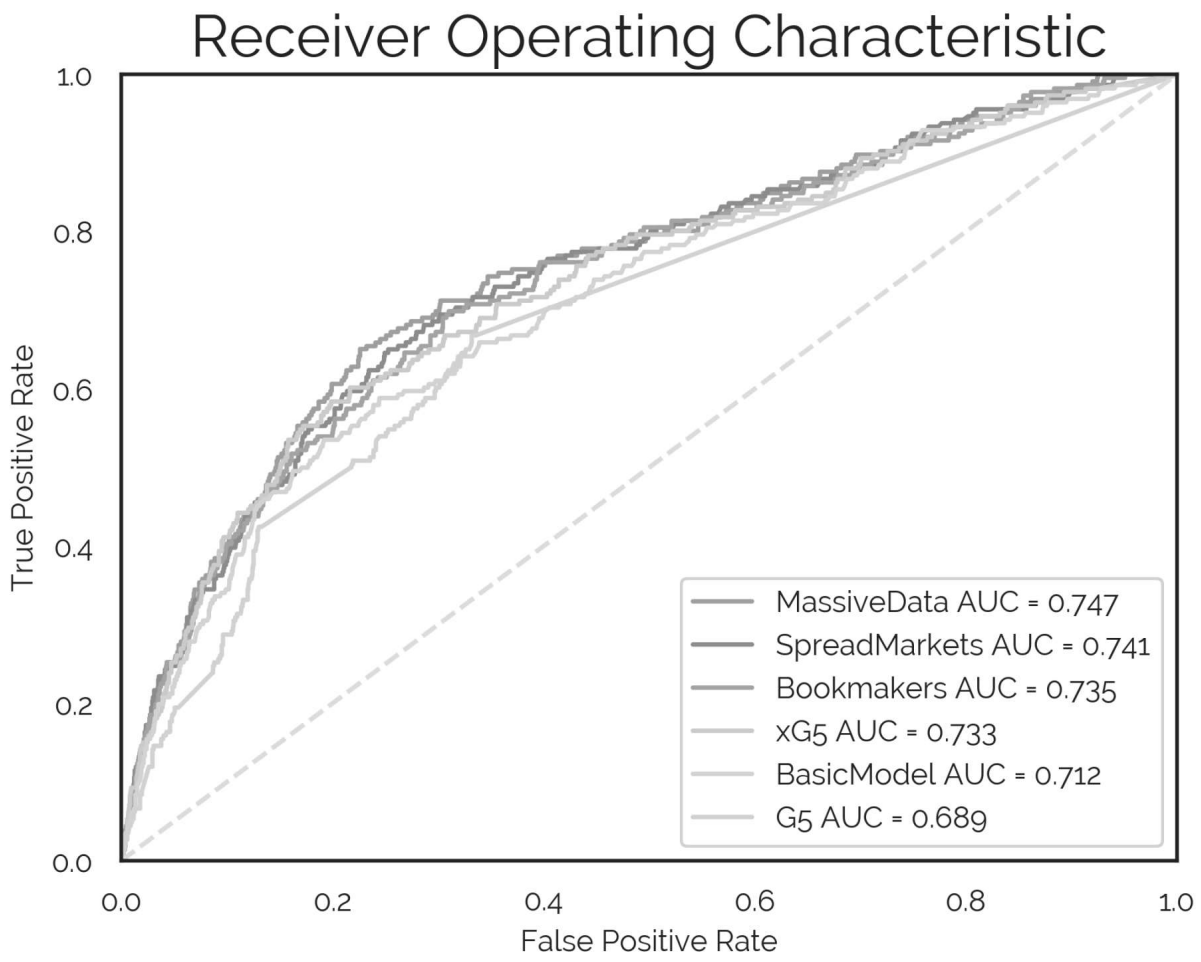


Anytime Percent & Hit Detection

Lets consider this analysis from a different point of view. A good model should also have the ability to indicate whether a player will hit (score any amount of goals) or not (score 0 goals). Conveniently this perspective reduces the target to an array of binary values. For this scenario the Receiver Operating Characteristic is a powerful tool (also used for analysing xG models which also have binary targets ie. goal or not).

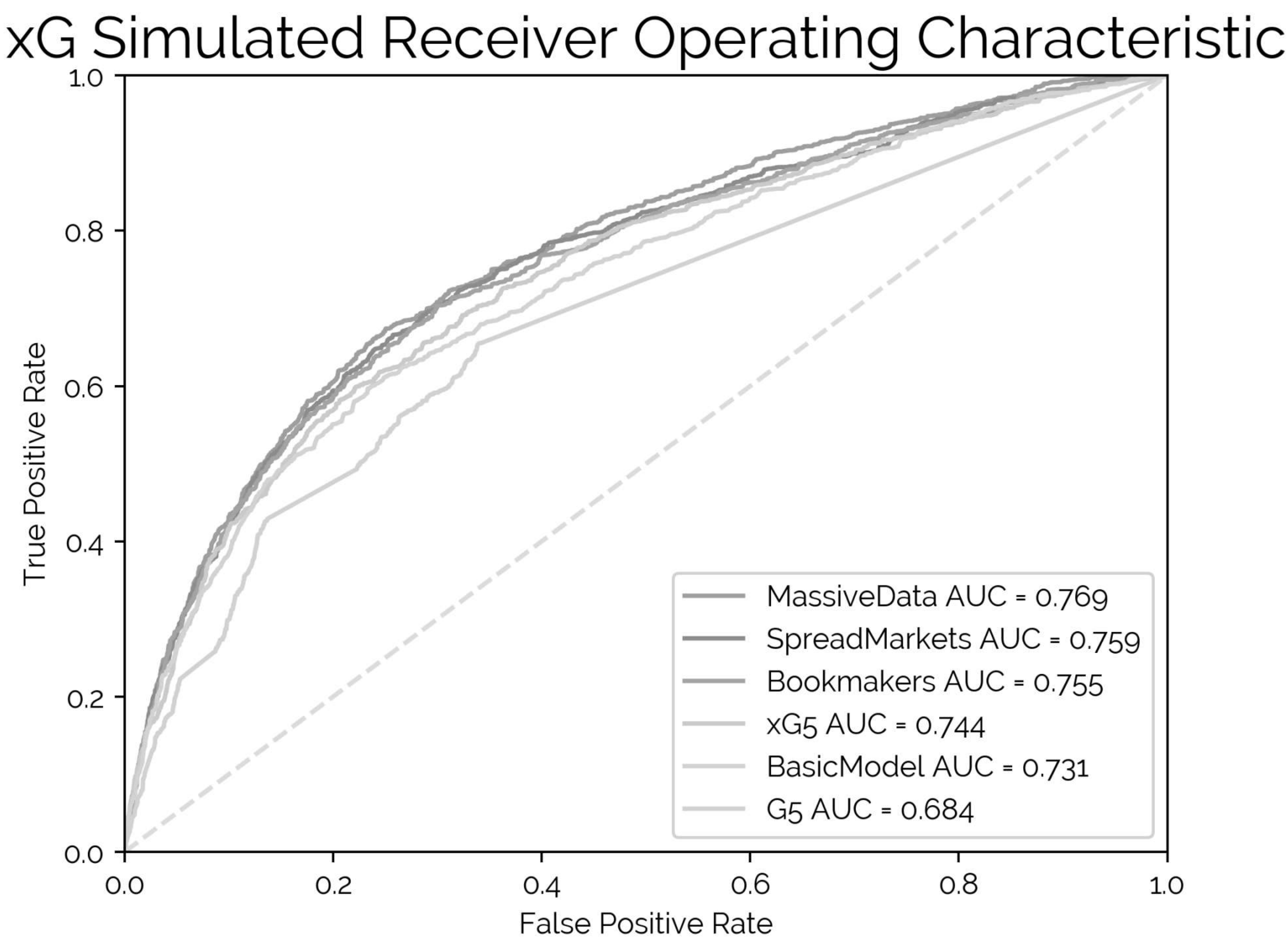
In terms of the predictive models, they simply need to be passed through a suitable distribution to convert their expected average amount of goals to an expected hit rate (better know as Anytime %). Once that is done we have an Anytime % available for each and every sample & model combination. fplreview.com actually releases this data for Implications Spread Odds and the Massive Data model.

The result we are interested in is the AUC (area under curve score), a high score indicates an ability to predict players who hit accurately, a low score indicates an inability to do so. The resultant plot and AUC scores can be seen below (the closer the curve to the top left, the better the predictive model).



ROC curve for the models anytime percent predictions

For fun I decided to create an xG based AUC curve- to achieve this I created a random goal generator based on the xG data and developed the 2,021 xG samples into 101,050 goals scored simulations. These goal simulations can in turn be translated into the binary hit/fail format required to generate a ROC curve. This creates a much smoother result and is probably the best visualisation of the performance of these models.



Higher definition ROC curve generated with xG data

As we can see above the trend found in the earlier analysis continues ie. MassiveData>Spreads>Bookies>xG>Basic>Goals.

Interpreting the Results

Across each of the analysis techniques above consistent trends have emerged, let's get an overall view of whats going on.

The Massive Data model has the best prediction power based on all of the above metrics and targets. Not far behind is Implications from Spread Odds (used as the source in the Implied Odds Team Planner). In third is Implications from Bookies Odds- they were comfortably beaten but still appear as a useful quick and easy source. The added value seen from Spread Odds is worth thinking about.

One point to get across even at the expense of you giving a moments notice to the fancy tools on my website that I'm trying to show off is how terrible of a predictor recent goals are.

The amount of goals a player has scored in his last 450 minutes is rendered almost worthless relative to the other indicators. Even the basic dummy model which is based on a players FPL position and ted fixture difficulty is far, far superior. That simple model would stupidly rate Fabinho as Salahs valent- it still beats focusing on recent goal data...

Recent xG (450 minutes of data) is marginally better than the dummy FPL position model. It would be fair to say that knowing a players real position and his fixture difficulty should give you a better read than recent xG. That said, change perspective by using longer-term xG data and it quickly becomes a useful metric and over ~20+ games can be used to create a model close in performance to bookies implications. xG data is designed to be descriptive- ie. provide context for a past event- this is markedly different to being predictive, especially with a small amount of data.

The Massive Data Model

Given that the Massive Data model is found to be so powerful, what exactly is it?

Well there are limits to what I will reveal but I can say it is a Machine Learning algorithm being fed with all kinds of performance data over various time-frames. The real key is contextualising past & future events to derive meaningful predictions from performance data and it get's back to the core of the quote at the start of this post. It's a big deal if a player is moved on/off penalties, if his position has changed, he has a history of being an excellent/terrible finisher or his team is about to play easier games etc. If a model is not thinking in that type of world it is going face limitations.

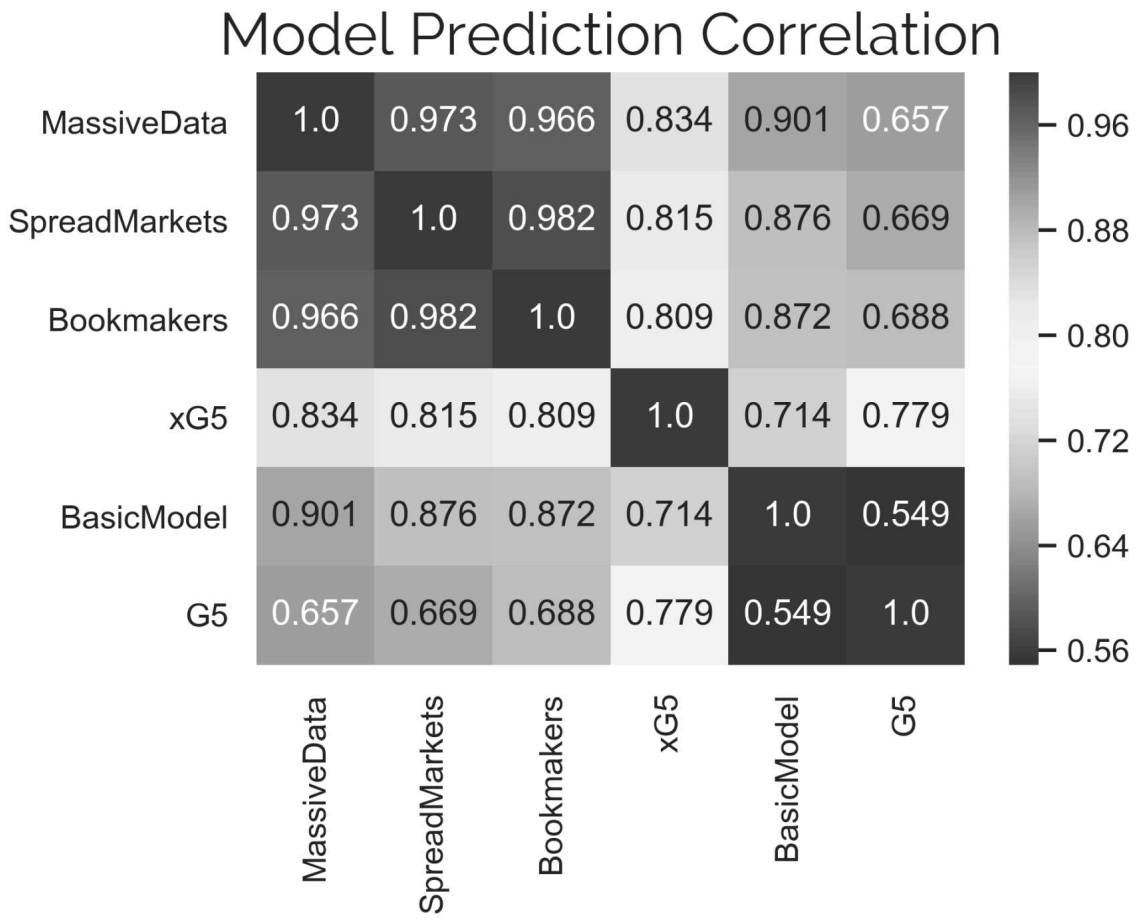
With enough meaningful context and background data it is possible to make very powerful forecasts that are more useful than trying to read through the murky waters of bookies margins on scorer odds.

On top of increased predictivity there are other (more) valuable gains:

- The Massive Data model can predict as far in the future as desired. A big limitation with Odds is that forecasts at player level are limited to 1 GW ahead and require the use of a deductive model to infer underlying beliefs for future predictions
- If the model knows Milner will return from injury in 2GWs, Salahs ownership of penalties will drop for those future projections- the same applies to expected future positional changes etc. The long term predictions are more dynamic
- The model is quite new and while it has been pushed very far there will be room for future refinement. In fact I already have ideas! That said each incremental gain becomes harder and harder so likely only minor wins left.
- On top of predicting goals, the model predicts FPL assists with exactly the same core concepts
- Backtesting was also performed for 2018/19 against bookmakers odds and the same result was repeated- the Massive Data comfortably beat bookies implications.
- This is not a case of fitting a model to suit this year. The model was learnt (train & test data) from data before the 19/20 season. The same was done for the 18/19 backtest. Data leakage would artificially inflate performance so I ensured that did not occur here.

While the downside of requiring enough data is minor. Should the model have less than 5 full games worth of data on a player it defers to Spread Odds- which are also a fantastic source!

It's important to stress to any gamblers reading this and hoping to take advantage, please do not think this is a money making model and I do not advise using it as such. While it is more predictive than implications from odds (ie. margins removed) you will just lose less money- the margins for goalscorer markets are simply too high in general (the average margin is a whopping 45% and bookies are very defensive with these odds in general). If there were millions to be made I would not share this tool.



Correlations between all models. Massive Data & Spread Markets leaning more towards recent xG than recent goal data relative to Bookmaker Markets

Notes on Implied Odds from Bookmakers/Spread Markets

Using anytime goalscorer odds from bookmakers is very popular and for good reason- many people use gambling sites and they really do a give a solid quick indicator of expected performance levels. However my data indicates Spread Markets scorer data to be more indicative and I'll run through a few reasons why that is the case.

Firstly what are Implied Odds from bookies? Well supposing Salah has odds of 2.50 to score at anytime in a match. That infers he has a 1/2.5 chance (40%) of scoring. Unfortunately many people stop there think you should expect him to earn 5*40%=2pts from goals. But this ignores bookmakers [illegible] margins to earn profits and also that Salah can indeed score more than 1 goal- after doing [illegible] two tasks you are left with an implication of how many goals the player will score on average.

Decoding bookmaker margin treatment in scorer markets is really a bit of a dark art. The common perception is that using the longest available odds will avail of a sharp bookmaker who is cutting it close to the true odds. This can happen but they are not necessarily consistent in doing this which creates a problem. Performing this action for all 500+ players still leaves a margin significantly greater than we would see in W/D/L or CS markets (which goes to show why this is a market to avoid putting money on). So now not only is there still margin remaining, it's not applied in a consistent manner.

A less popular approach is using a standard margin cleaning method on all available odds and getting the mean expectation per player. From my analysis this performs very marginally better but it also falls into traps- bookies are often cautious of certain types of situations, some bookies just follow the trend and of course they will all look to take advantage of the same money earning tactics. This means correlated distortions will feed into the model, which wouldn't necessarily happen with the other approach. On the upside much greater consistency with margin treatment is achieved- which is valuable. Overall the differences are marginal and for the sake of this analysis I have gone with the better performing model.

So the with the two pros/cons of the best approaches to bookmaker odds considered why would the less popular Spread Markets be any better? To put it simply they have much less space to hide- Spreads are two sided opposing markets. Margin treatment is not necessarily as easy as it may first seem but it's a lot more solvable. The nature of this market makes it significantly more informative than bookmaker markets as we can see pessimistic and optimistic reads per player- it should be no a surprise how well it performed in testing. These markets feed the implied odds data in the Team Planner- Implied Odds tool and per 90 values are available here.

Conclusion

Overall it appears implications bookmakers odds (at player level) have been comfortably dethroned. The Massive Data model has outperformed it with ease in 2018/19 and 2019/2020 seasons. Intuitively Spread Odds are also more powerful based on 2019/2020 data. More data will be needed to confirm whether the Massive Data model definitely outperforms implications from Spread Markets but the initial data is warm on that idea. The other gains on top of potential predictive power (ie. Spreads have player-level data generally ~2 days before kick-off, the Massive Data model will have that data available months in advance) would give even greater confidence in the Massive Data model.

The easiest result for fplreview.com would have been for bookmaker odds to come out on top which would have left the site as designed per the 2018/19 season, however analysis like this drove the Implied Odds Team Planner to move to Spread Odds and the creation of the Massive Data Team Planner to run in parallel.

ns of those determined to make predictions purely from xG data, the focus should be on longer data than 5 matches. Simply knowing a players actual role and upcoming fixtures will likely be

more valuable until enough data is considered.

Recent goal data gets far more energy and thought than it deserves- it very often acts as a red herring and leads to managers chasing points and switching players needlessly.

← Previous Post

Next Post →

About
Privacy Policy
Cookie Policy
Terms of Service

Contact
formdork@fplreview.com