Big 6 vs. Everybody: An Expected Goals Analysis of English Premier League Strikers

Kien Vu '21 , Dickinson College

Introduction:

Despite being arguably the most popular sport in the world, global soccer has been reluctant to adopt analytics and especially advanced metrics. However, clubs and organizations are increasingly utilizing modern tracking systems in order to get ahead in scouting and tactical understanding. One metric that has been attracting public attention is expected goals, which aims to estimate the number of goals scored in a game by assessing the situation and traits of the shots launched. A team wins by scoring more goals than the other, thus the ability to understand goals would be invaluable in the decision-making process. I hope to examine expected goals with a focus on the strikers, in order to understand how a player's tactical role and their playing environment affect team goal output, more specifically passing, playmaking, shooting, and team passing traits.

The English Premier League is considered one of the best and most competitive leagues in the world, with the highest revenue of any sports league outside America. While a sizable portion of the cash is shared equally between the clubs, performance parity is often not the case. In fact, in the last three seasons the same clubs (Arsenal, Chelsea, Liverpool, Manchester United, Manchester City, Tottenham) occupy the top 6 positions in the standings, winning and scoring more by a wide margin. These teams' success on the pitch is clear in the results, and I hypothesize that there are factors in how their strikers play that support and also reflect the superiority of these teams.

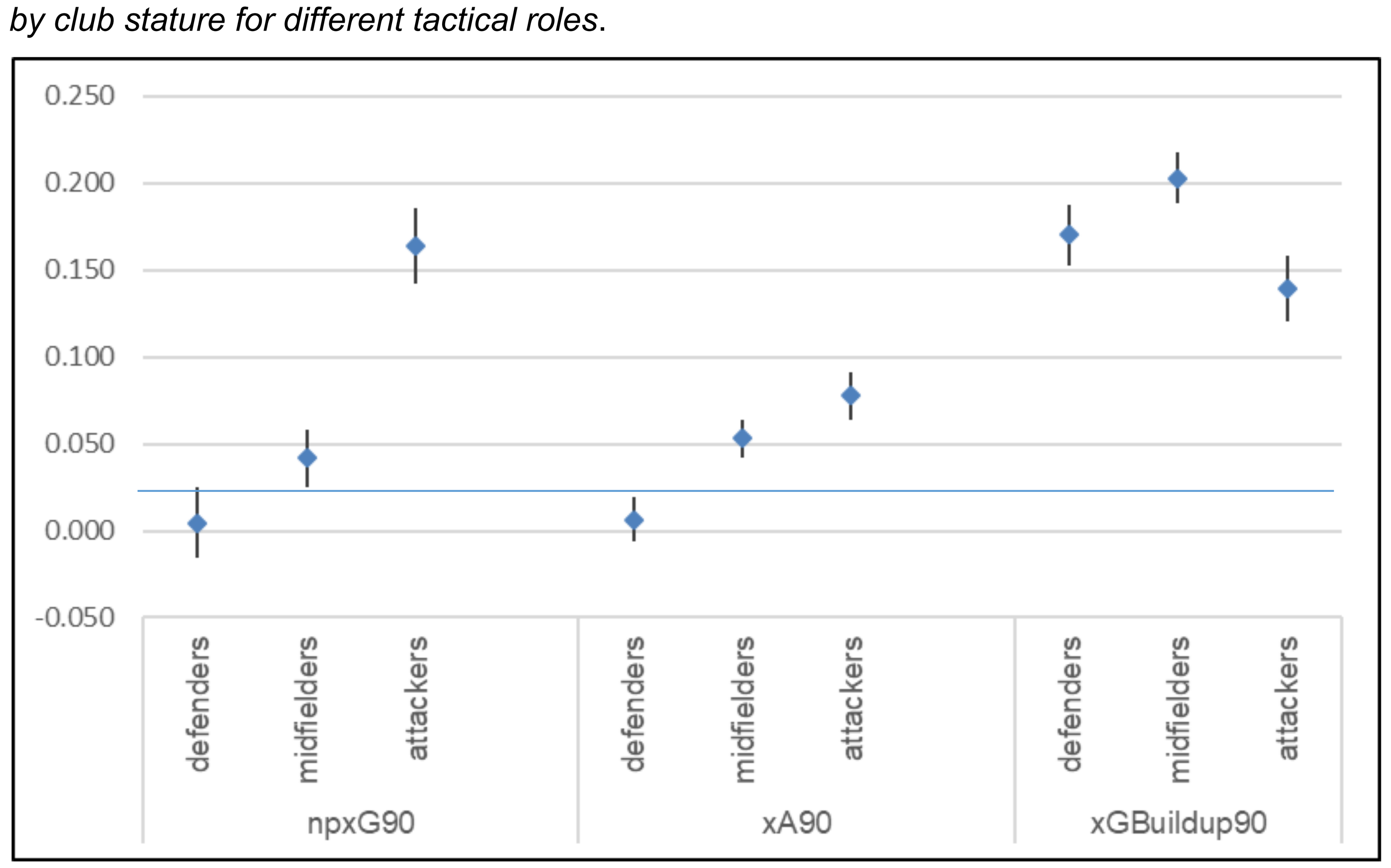
Methodology:

The full dataset was gathered from understat.com, updated on March 27th , 2019. The data includes 2613 individual player statistics and team statistics from the 2014-2015 Premier League season to the current 2018-2019 season. SPSS statistical software was used to run ANOVA post hoc tests, multivariate regressions, and bivariate Pearson correlation.

The player offensive metrics analyzed here includes expected goals from non-penalty shots (npxG90), expected goals from key passes (xA90), and expected goals from build up play – possessions involved without key passes or shots (xGBuildup90) for each player, adjusted for minutes played. One-way ANOVA analysis was used to find the standardized mean difference in offensive metrics between non-goalkeeping big 6 players and the others, in subgroups according to position.

Multivariate regressions was performed with each of the expected goals variables above as dependent variables, but with the scope narrowed down to strikers. Dummy independent variables includes those about whether the player is in a big 6 team and about the team's offensive style and quality, which sorts out greater than median values of Average number of passes completed in the opposition half and average of passes completed within 20 meters of the goal. The minutes per game statistics of the player is also included as an independent variable. Bivariate Pearson correlation analysis was performed for each pair of variables.

Figure 1. *Expected Goals from shots, key passes and buildup: Standardized mean difference by club stature for different tactical roles.*



Regression

Table 1.
Multiple Regression Analysis of Expected Goals on Three Types of Involvement

Variable	xA per 90	xG per 90 Model 1	xG per 90 Model 2	xG from Buildup per 90 Model 1	xG from Buildup per 90 Model 2
Constant	-0.01 (0.02)	0.08 (0.04)	0.15 *** (0.04)	0.05 *** (0.02)	0.08 *** (0.02)
xG per 90	-0.02 (0.02)			-0.03 (0.02)	-0.08 *** (0.02)
xA per 90		-0.11 (0.13)	-0.10 (0.12)	0.58 *** (0.05)	0.51 *** (0.05)
xG from Buildup per 90	0.43 *** (0.04)	-0.14 (0.11)	-0.38 *** (0.11)		
More pass completions	0.00 (0.01)	0.06 *** (0.02)	0.00 (0.02)	0.04 *** (0.01)	0.01 (0.01)
More deep completions	0.02 * (0.01)	0.10 *** (0.02)	0.04 (0.02)	0.05 *** (0.01)	0.02 * (0.01)
Minutes per Game/10	0.01 *** (0.003)	0.02 *** (0.01)	0.02 *** (0.01)	0.00 (0.003)	0.00 (0.003)
Big Club	0.00 (0.01)		0.19 *** (0.03)		0.01 *** (0.01)
F	50.09	14.08	20.91	80.03	85.59
Adjusted R^2	.42	.14	.22	.49	.55

Note. For all five regressions, *N* = 428. Raw regression coefficients (with standard errors beneath each coefficient in parentheses). xG = Expected Goals. xA = Expected Assists.
$^{*}p < .05$. $^{**}p < .01$. $^{***}p < .001$.

Table 2.
Correlation Matrix

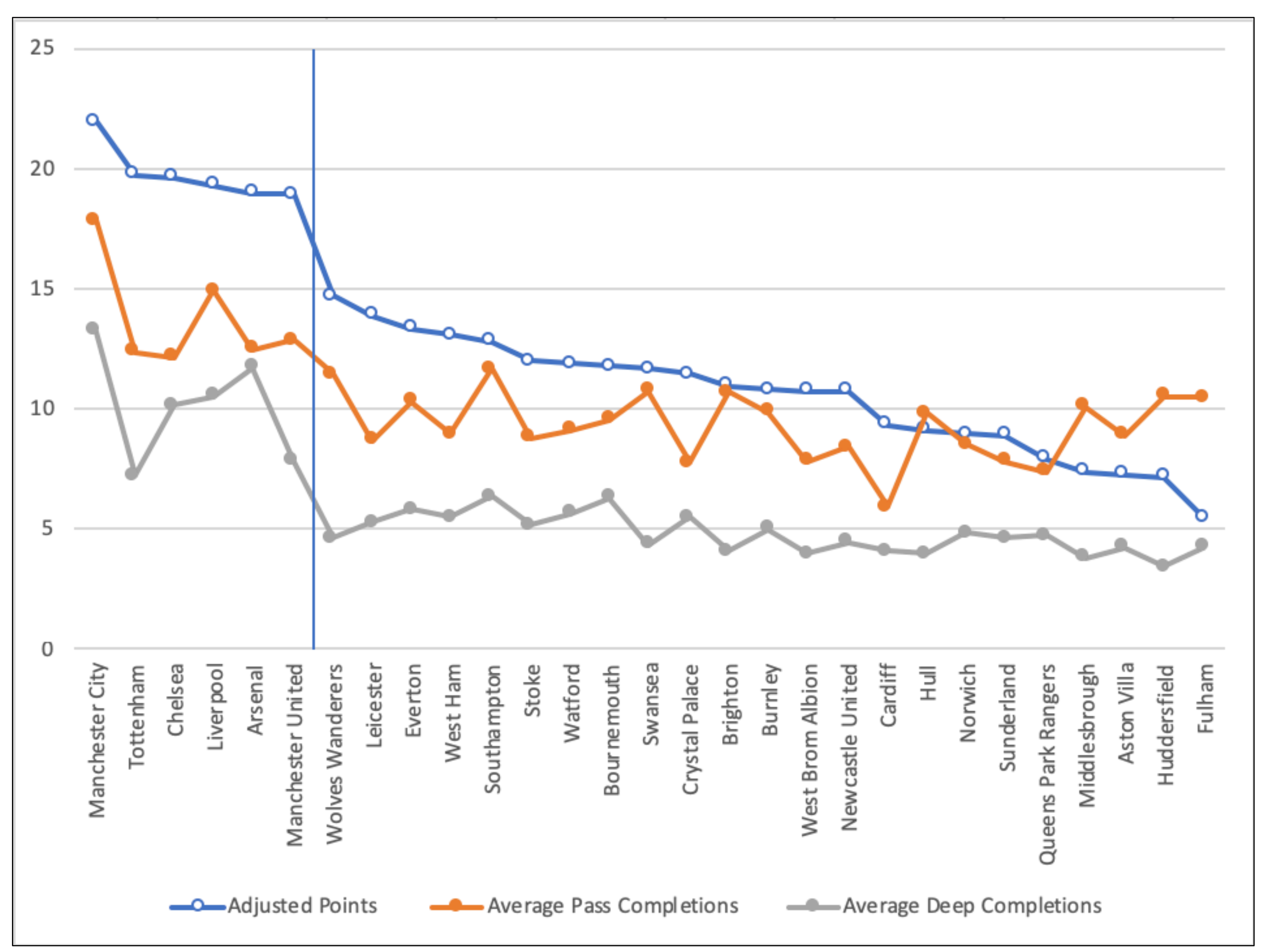
	npxG90	xA90	xGBuildup90	gtmOPPDA	gtmDC	bigclub	minpgame10
npxG90		.030	.011	p< .001	p< .001	p< .001	p< .001
xA90	.105*		p< .001	p< .001	p< .001	p< .001	p< .001
xGBuildup90	.123*	.629**		p< .001	p< .001	p< .001	.009
gtmOPPDA	.230**	.292**	.420**		p< .001	p< .001	.485
gtmDC	.315**	.405**	.497**	.368**		p< .001	.006
bigclub	.417**	.413**	.603**	.600**	.636**		.006
minpgame10	.180**	.238**	.126**	.034	.133**	.133**	

Note: Lower triangle indicates the correlation coefficients. Upper triangle indicates the two-tailed p-values.
*. Correlation is significant at the 0.05 level (2-tailed).
**. Correlation is significant at the 0.01 level (2-tailed).

Table 3
Variables Key

xG per 90 (npxG90)	Non-penalty Expected Goals per 90 minutes.
xA per 90 (xA90)	Expected goals from a key pass, per 90 minutes.
xG from Buildup per 90 (xGBuildup90)	Expected goals from a non-shot-or-key-pass involvement, per 90 minutes.
More pass completions (gtmOPPDA)	Whether one's team is above median in passes completed per possession in opposition half.
More deep completions (gtmDC)	Whether one's team is above median in passes completed per game within 20 meters of goal.
Big Club (bigclub)	Whether one's team is a Big 6 club.
Minutes per game/10 (minpgame10)	Average minutes per game played.

Figure 2. *Adjusted Points, Average pass completions and Average deep completions per game for Premier League teams from 2014-2015 to 2018-2019. (Note: Big 6 clubs are left of the vertical line).*



Regression results and analysis:

- Looking at expected goals scored as dependent variable in Table 1,
  - First model provides strong evidence ($p < .001$) to suggest a positive correlation with minutes per game, and the striker's team completing more passes in opposition half and in dangerous areas to goal-scoring. However, its explanatory power is low ($R^2 = .14$).
  - Rather, being in a big club is suggested to increase shooting expected goals, and having higher expected goals from buildup decreases it. The two passing dummy variables are also no longer positively correlated with expected goals.
  - ❖ This implies that playing in a more passing heavy team and making more dangerous passes doesn't mean scoring more if a striker is in a same–tier club.
  - ❖ Moreover, strikers who involve more in build up are less likely to score goals when club stature is controlled.
- For expected goals from buildup and assists as dependent variables in Table 1,
  - When controlling for big club, having higher expected assists and being in a big team correlates positively to buildup success. The model has a stronger explanatory power ($R^2 = .55$)
  - With the assist variable, being in a big club does not imply a change, but being involved in buildup is correlated to have a positive relationship.
  - The team passing variables decrease in statistical significance in the second buildup model as is with the scoring model. Deep completions have a lower coefficient, but still significant ($p < .05$) relationship to both buildup and assist variables.
  - ❖ Strikers who assists more are likely to involve in buildup of more goals, due to the strong positive relationship between two variables. These players are also likely to be in teams that completes more passes near goal.

Variable correlation analysis(Table 2) does not need multicollinearity considerations.

Discussion:

For the many ways of a player to contribute to a goal in soccer, a striker is most often associated with scoring. Accordingly, the results infers that most strikers might have lighter buildup duties to focus on scoring. However, those who involve in more buildups are also more likely to assist, meaning that there are strikers whose duties extend to playmaking tasks as well. For big clubs, a significant, positive relationship means that the strikers from these club are likely to be involved in both scoring and buildup.

Even with the responsibilities, big club strikers score in a much greater rate, demonstrated by both the standardized mean difference in Figure 1 and the regression. Not only are strikers more well rounded, the ANOVA analysis in Figure 1 show midfielders and defenders being involved in more successful possessions as well. Having players who can be involved in more phases of play in a more effective way can definitely contribute to greater success.

While the players themselves might be responsible for the output, other factors such as the style of play might also be important. From the regression, notice the change in significance of team passing variables, which indicates propensity for passing and deep completions, when big club is controlled. In addition, the correlation analysis in Table 2 shows high correlations for big club dummy with the two variables. Thus it is likely that these big teams often employs more passing-driven style and play in a more intricate, indirect manner. Figure 2 shows the three variables for each team, and there is a clear difference for the big 6 teams and the rest.

However, the correlation between this specific style and greater scoring or success is not apparent. When controlled for club stature, both pass completion and average deep completion dummy variables have a statistically insignificant coefficient to expected goals. Therefore, for other teams, applying these tactics does not mean that goal-scoring would increase. Coupled with the relatively low explanatory power of the expected goals scored regression ($R^2 = .22$), the analysis is inadequate to explain more of expected goal-scoring.

Conclusion:

- More goal-friendly strikers are less likely to create good buildup opportunities.
- Strikers from big 6 teams tends to involve in both tasks more, and better.
- The more involved passing style of good teams are not necessarily the key to success.