1-1-2018

# Building a Soccer Dynasty

Carter Bedsole
cbedsole15@georgefox.edu

Gary Buckley
gbuckley15@georgefox.edu

# Building a Soccer Dynasty

Carter Bedsole and Gary Buckley
George Fox University

Abstract

The soccer world operates as a free market. Buying, selling, and trading of players is vital to the success of a club. A successful soccer club brings a great deal of revenue and growth to a local economy. Therefore, clubs ought to be prudent when signing players. In this paper, we use ordinary least squares regressions on Major League Soccer player data from 2015-2018 to determine the effect strikers and goalkeepers have on team success. In other words, what is the marginal impact of a good striker relative to a bad one? A good goalkeeper relative to a mediocre one? Finally, we include salary data to determine if clubs are paying strikers and goalkeepers according to their performance and production over the course of a season.

Keywords: Major League Soccer, Econometrics, Striker, Goalkeeper, Salary, Performance.

JEL Codes: Z200, C510, C530.

## 1. Introduction

Any good business seeks a lucrative return on investment. World soccer is no different. Most clubs operate in a free market buying, selling, and trading players as finances allow. The return on the investment of players is wins. General Managers in Major League Soccer (MLS) have a unique challenge: working with a limited budget equal to that of all other clubs in the league. In this context, maximizing player production on a budget is paramount to having a successful team.

Not only is player production important for the success of a team, but it is also important for the vitality of a city. For example, two years ago the city of Atlanta did not have a top flight soccer team. Through savvy player investments, the Atlanta United management built a championship team by 2018. Because of their success, Atlanta United's attendance was in the top 20 for soccer clubs worldwide (Bogert, 2018). In the season finale, Atlanta hosted the MLS Cup Final, drawing well over 70,000 fans and generating a great deal of revenue for the city. The positive impact of a strong fan base on the local economy is undeniable.

How does a club replicate the success Atlanta United has seen? How does a club build a successful team that pleases both fans and the local economy? Soccer is a team sport. As such, the sport wrestles with the tension between acquiring talented individual players and forming a talented team. While team chemistry is essential to success, it often cannot be bought. However, talented, productive players *can* be bought. Thus, clubs in MLS ought to acquire talent prudently.

In this project, we seek to answer two questions about building a successful soccer team. First, how do talented goalkeepers and strikers (also referred to as forwards) affect their teams' performances over the course of a season? We build a regression model to estimate the effect of player production in those positions on a team's record. Second, we seek to determine if goalkeepers and strikers are paid appropriately based on their production. We build secondary regression models to determine if player production has a positive effect on player salary.

We choose to analyze goalkeepers and strikers due to the relative simplicity and availability of statistics. With more advanced and more extensive data, the same methodology could be applied to other positions.

Properly analyzing the impact of an individual soccer player is challenging. In tackle football, a quarterback’s yards, completion rate, touchdowns, and interceptions are often enough to tell the story of his game. However, in soccer a player could have a spectacular game that does not show in the game’s statistics. A goalkeeper could not allow any goals but not face any shots. A striker could generate good chances for his teammates but be unable to score. The scoreline and match stats sometimes fail to tell a complete story of how a player, and even an entire team, perform. Therefore, advanced analytics must be used to fill in the missing pieces of the story. American Soccer Analysis (ASA), a data-driven soccer blog, records advanced statistics for MLS. Using their data for Expected Goals For and Expected Goals Against statistics, we can distill individual player performance. ASA also provides more traditional statistics such as shots (for a striker) and shots faced (for a goalkeeper). By combining advanced and traditional statistics, we create a measure for player production over a season. Finally, we use ASA’s MLS salary data, along with extra salary information from the MLS website, to discuss a player’s value.

Our models rely on Ordinary Least Squares (OLS) regressions of a team’s total points (performance) on a number of explanatory variables, including goalkeeper and striker performance. Our findings demonstrate that a goalkeeper’s performance has no statistically significant effect on a team’s success over the course of a season. A striker’s production does have a statistically significant effect on a team's success, but the economic significance of this effect is questionable. Additionally, we determined that strikers are indeed paid according to merit. Our findings demonstrate that goalkeepers are not paid according to merit, but perhaps this could be teased out a bit more with more data or a different approach.

In this paper, we begin with a detailed discussion of the data and the assumptions we are making with it. Additionally, we include explanations of soccer concepts and argue for the significance of the data chosen. In the following section, we present our theoretical models for our research questions and state our hypotheses. Finally, we present and analyze our results, and seek to answer our two research questions.

## 2. Data Overview

### 2.1 Data Source and Expected Goals Explanation

We retrieved our data from American Soccer Analysis, a site that provides data-driven journalism on MLS and tracks player production and salary statistics. Our sample contains data from four complete MLS regular seasons (2015 - 2018).

Our most important data from American Soccer Analysis is expected goal data. Soccer is a relatively low-scoring sport, and so often the scoreline does not accurately represent the performance of a team or individual. For instance, a team could take 50 shots and still lose, while another team could take a single shot and win. ASA's "Expected" model attempts to mitigate random noise and provide an accurate picture of a team's performance (*What Are Expected Goals?,* 2017). Using historical MLS data, this model analyzes the likelihood of each shot to result in a goal and sums these likelihoods to create the "expected goals" metric.


*Figure 1*: The expected goals map of the November 29th, 2018 match between Sporting Kansas City and Portland Timbers. This map uses circles to represent every shot taken in the match. Yellow circles represent shots that resulted in goals and blue circles represent shots that did not result in goals. The size of a circle represents the likelihood of that shot to result in a goal. The expected goals model in this figure was provided by Opta, but it is in practice very similar to the model we have chosen to use from American Soccer Analysis. Taken from: BenBaer89. (2018, November 29). #SKCvsPOR xG. Tough one for SKC [Tweet]. Retrieved from https://twitter.com/BenBaer89/status/1068379531445178368

A good example of the power of the expected goals model is the November 29th, 2018 match between Sporting Kansas City and Portland Timbers, shown in Figure 1. Portland Timbers won the match 3-2. However, as shown by the expected goals model, perhaps they got lucky. Sporting Kansas City generated more shots and better shots and would have been expected to win the match approximately 3-1, as given by the Total Team xG in the figure. This is why we turn to expected goals: to paint a more complete picture of a team's performance than the scoreline may indicate. (This same methodology will be applied to capture the quality of an individual's performance as well).

**2.2 Data Assumptions**

Our sample data comes from the 2015-2018 MLS seasons. In this context, we will ignore the effects of inflation on player salaries. Since contracts are often negotiated over multiple years we expect player salaries to remain relatively stable in our small sample. Additionally, since a salary cap stifles normal wage growth, it is justified to ignore the effects of inflation. Similarly, we will ignore the impact of rule changes to the salary cap on player salaries. The league has instituted some salary structure changes in the time period we are studying; however, since we are studying the relationship between salaries and production, not the dollar amounts themselves, these structural changes will not affect our research.

We are assuming this data is cross-sectional. Because our sample covers a relatively small time range, we do not expect to see significant time trends. We will treat each season's player and team statistics as independent. We view the performance of a player in a given year within our four-year period to be a random sample of their true ability, rather than a reflection of their true growth or decline. Therefore, since we view each year as a random sample of a player's ability and each year as a random sample of the league's norms, we will interpret the data as cross-sectional.

Finally, we believe that there is little omitted variable bias in our research. As explained, American Soccer Analysis has already accounted for much of the random data noise in their expected model. We believe that this expected model accurately captures team and individual performance.

# 3. Methodology

## 3.1 Theoretical Model

### 3.1.1 Primary Model.

Our primary regression model will take the form:

$$Points = \beta_0 + \beta_1 * xGF + \beta_2 * GSA\text{-}xGSA + \beta_3 * xGA + \beta_4 * GA\text{-}xGA$$

Explanations and interpretations of the variables are as follows:

***Points.***

This variable captures a team's record over a full season. In soccer, a team is awarded 3 points for a win, 1 for a draw, and 0 for a loss. The sum of all points is indicative of a team's performance throughout a season. Our model seeks to determine which factors influence a team's ability to earn more points.

***xGF (Expected Goals Scored).***

This variable represents the total expected goals a team created during a season. Similar to how American Soccer Analysis sums each shot's likelihood to result in a goal to generate expected goals in a game, this variable sums the likelihood for all shots in a season. A higher xGF value is indicative of a good offense, which produces a lot of shots that should turn into goals.

***GSA-xGSA (Actual Goals and Assists - Expected Goals and Assists)***.

The GSA portion of this variable is the sum of goals and assists that a striker recorded during the season. The xGSA portion is the expected goals and assists. Like with shots and goals, American Soccer Analysis uses a model to predict how likely each pass is to be an assist. Summing expected goals and assists gives us xGSA. By subtracting xGSA from GSA, we distill the quality of an individual player. If the player has a positive value, he is being more productive than the average player would be in his exact situation. If the player has a negative value, he is being less productive than the average player would be in his situation. In our sample, we only consider strikers who have played at least half the season (and have thus had a significant impact on a team's results). We will consider this variable for both primary strikers (most minutes played on the team) and secondary (second most minutes played on the team).

***xGA (Expected Goals Against)***.

xGA measures the total expected goals against during a team's season. Similar to xGF which measures the likelihood a shot is scored, xGA measures the likelihood that the opposing team's shot is scored. A low value of xGA represents a good defense which does not allow opponents to take dangerous shots.

***GA-xGA (Actual Goals Against - Expected Goals Against).***

GA represents how many actual goals a goalkeeper conceded. xGA is the expected goals against based on the quality of shots the opponents took. A positive value is indicative of a bad goalkeeper who let in more goals than the average goalkeeper would have in his exact situation. A negative value is indicative of a good goalkeeper who conceded fewer goals than expected.

We are measuring this variable for goalkeepers that played for at least half of the season and thus had a significant contribution to the team's performance.

In summary, we expect points to be a function of four variables (1. Quality of team's offense [*xGF*], 2. Quality of team's individual strikers [*GSA* - *xGSA*], 3. Quality of a team's defense [*xGA*], and 4. Quality of a team's goalkeeper [*GA* - *xGA*]). (We have verified that these explanatory variables are not perfectly collinear). Our regression determines the effect of a team's individual strikers and goalkeeper on overall performance, while controlling for the quality of a team's attack and defense.

***Summary Statistics***

Table 1 includes summary statistics for the explanatory variables in our model.

| ***Table 1*** | | | | | | | |
|---|---|---|---|---|---|---|---|
| *Summary Statistics of Explanatory Variables* | | | | | | | |
| Variable | xGF | GSA-xGSA (Primary Forward) | GSA-xGSA (Secondary Forward) | xGA | GA-xGA | Forward Salary | Goalkeeper Salary |
| Mean | 45.501 | 0.618 | 0.680 | 45.499 | -0.778 | $988,610 | $263,660 |
| Median | 44.800 | 0.040 | 0.000 | 45.100 | -0.420 | $432,500 | $157,920 |
| Minimum | 34.400 | -6.050 | -3.810 | 31.700 | -13.310 | $53,472 | $65,625 |
| Maximum | 65.800 | 8.400 | 9.330 | 63.300 | 8.390 | $5,610,000 | $2,100,000 |
| Standard Deviation | 6.543 | 3.134 | 2.085 | 7.048 | 3.995 | $1,462,500 | $368,550 |

Table 1

*xGF*: Expected Goals Scored.
*GSA* - *xGSA:* Actual Goals and Assists - Expected Goals and Assists.
*xGA*: Expected Goals Against.
*GA* - *xGA*: Actual Goals Against - Expected Goals Against.
Note: this sample includes 74 observations.

### 3.1.2 Secondary Model.

We will pursue a secondary research question in light of the first: are goalkeepers and strikers paid according to their performance and value? To do so, we will use the following models:

$$F\_Salary = \beta_0 + \beta_1 * GSA\text{-}xGSA + \beta_2 * xGSA$$

$$GK\_Salary = \beta_0 + \beta_1 * GA\text{-}xGA + \beta_2 * GA$$

In these models we distinguish between performance and production. For strikers, we define performance as GSA-xGSA. Performance measures how well a forward did compared to how well the average forward would do in their situation. Production is a striker's xGSA - how many expected goals and assists he actually created. Production measures how much of an impact a striker had. Similarly, goalkeeper performance is measured by GA-xGA, which indicates if a goalkeeper performed above or below average. Production is measured by GA and indicates how many goals a goalkeeper actually conceded (we use GA instead of xGA since a goalkeeper has little impact on if the opposing team takes a shot but has a big impact on if that shot turns into a goal).

With these regressions, we seek to discover if more talented goalkeepers - measured by performance and production - have higher salaries.

## 3.2 Hypotheses

We have three primary hypotheses for our research. First, we expect goalkeepers to be the most important influencer of points. By most important, we mean that we expect all four explanatory variables in the primary model to have statistical and economic significance;

however, we expect GA-xGA (goalkeeper performance) to have the largest coefficient value and thus be the most important influencer on points.

Second, in contrast to the first hypothesis, we expect goalkeepers will not be paid according to merit. Thus, goalkeeper salary will not be correlated with GA-xGA or GA, implying that teams do not accurately compensate their goalkeepers.

Finally, we think that strikers will be paid according to merit. Since a striker's contribution to a team's success is easily seen through highlight goals and assists, we expect that more productive strikers will be receive higher salaries.

## 4. Results

### 4.1 Season-long Success

#### 4.1.1 Overall Offense and Defense

Table 2 summarizes the impact of overall offense and defense and the impact of individual striker and goalkeeper performance on points. In this sample, we only consider teams that had a goalkeeper who played more than half of the available minutes in a season. We also only consider teams with at least one forward who played more than half of the minutes available in a season. After these restrictions, we have 74 observations.

In Model 2.5 of Table 2, both xGF and xGA are statistically significant at the 1% level, suggesting that overall team offense and defense affect points. In line with expectation, xGF has a positive impact on points; generating better chances on offense leads to more points. For every expected goal that a team creates over the course of a season, they can expect to earn, on average, .59 more points. The range of xGF in our sample is 31.4. Our model predicts that teams with very good offenses that produce high xGF values can earn up to 18 more points than teams with poor offenses when all else is equal. Since the average points in our sample is 47.8, this is a gigantic impact. Also in line with expectation, xGA has a negative impact on points; the more good chances that a team concedes the fewer points they earn. For every expected goal a team concedes, on average, they earn .67 fewer points. xGA has a range of 31.6 so our model suggests that teams with good defenses can earn up to 21 more points than teams with bad defenses. Again, good overall defense is extremely important.

This confirms an old sports adage: defense wins championships. Although good offense and good defense both are important to team success, defense is more important: the absolute value of the coefficient on xGA is higher than that on xGF. This result is surprising since soccer theory suggests that it is much harder to score than to defend, suggesting an expected goal created would be more valuable. However, our results contradict this notion.

#### 4.1.2 Individual Forward Performance

How a team's forwards perform also impacts points, but the exact relationship is unclear from our results. In Model 2.5 of Table 2, the performance of a team's primary forward (the forward who played the most minutes of any forward on the team and played at least half the available season minutes) is statistically significant at the 5% level with a coefficient of .44. The best primary forward in our sample has a GSA - xGSA value of 8.4. All else being equal, a high-performing primary forward will earn a team 2 to 4 more points throughout the season, or approximately one extra win. In a season with only 34 games, one additional win is significant to a team.

However, when we include a team's secondary forward (the forward who played the second-most minutes of any forward on the team and played at least half the available season minutes) along with the primary forward, the relevance of a primary forward is questioned. Table

3 summarizes these results. Here, we restrict our sample further to teams that had two forwards that played over half of the minutes in a season. With these restrictions, we have 37 observations. While the impact of a primary forward's performance is still positive it is no longer statistically significant (Model 2.1). Surprisingly, when the performance of the secondary forward is added to the model (Model 2.2), his performance is statistically significant at the 1% level. This model suggests that a high-performing second striker (the maximum GSA - xGSA for secondary forwards is 9.3) can earn a team 6 to 8 more points over the course of a season.

This result is surprising. It seems that the performance of both the primary and secondary forward ought to impact points, not just one. A possible explanation is that the primary forward, since he plays the most, is likely the best forward on the team. He probably attracts the majority of attention from the opposing defense. A secondary forward will have more space and possibly more chances to score. Thus, a talented secondary forward can very valuable.

Another explanation is that the performances of both forwards matter, but are not as important as simply having more forwards. Table 4 shows the regression of points on number of forwards. Teams in our sample have either 0, 1, 2 or 3 forwards who played more than half the minutes of a season. The more forwards a team has, the more points they earn. Number of forwards is statistically significant at the 5% level. Its coefficient of 2.66 suggests that for every forward a team plays for more than half of the season, they will win about 1 more game. Teams with more forwards playing for more than half the season probably suffered fewer injuries, so it makes sense that they perform better. Further, it is likely that teams with more forwards are not tinkering with their lineup, so they have an established, successful style which often leads to more points.

The results indicate that the performance of individual forwards do marginally impact team success. But more important than any one player is a team's stability.

### 4.1.3 Goalkeeper Performance

Goalkeeper performance is not statistically significant (Table 2). This result initially seems surprising, but can be explained by considering the scope of the data. Our data considers season-long success. Though we do not have the data necessary to analyze this, it is clear that a goalkeeper's individual performance can significantly impact the outcome of a single game. But over a whole season, a goalkeeper's impact is small compared to total team defense. A team with a poor defense concedes about 60 expected goals (the 95th percentile of xGA is 60.0). An excellent goalkeeper will make up for only about 6 of those goals (the 95th percentile of GA-xGA is 5.94). In the short run, a goalkeeper can greatly impact if a team wins or loses, but in the long run, his performance has little effect on the number of points a team earns in a season.

## 4.2 Salaries

### 4.2.1 Forward Salaries

Table 5 summarizes our results analyzing forward performance and salary. In this sample, we do not consider the distinction between primary and secondary forwards. Instead, we analyze data for all forwards that have played more than half of the available season minutes in a season.

Both performance and production positively and significantly impact a forward's salary. A forward's performance (GSA - xGSA) is significant at the 5% level and a forward's production (xGSA) is significant at the 1% level. The coefficients ($91,320 and $115,020, respectively) are large and very significant to player salary. The coefficient on production is larger than that of performance and the range of production (29) is larger than the range of performance (16). Thus, a forward's production has a greater impact on his salary than his performance, though both significantly affect salary.

#### 4.2.2 Goalkeeper Salaries

Neither a goalkeeper's performance nor production impacts their salary. Table 6 summarizes our results. Both performance and production are statistically insignificant variables. Again, this result is surprising. Though our prior research suggests that goalkeeper performance does not affect long-term team success, we would still expect teams to pay goalkeepers in accordance with their performance or production. However, this is not the case. Further research is required to determine in goalkeeper pay is determined by other variables, perhaps something like highlight saves, or if it is unexplained.

## 5. Summary

Our data suggests that overall team offense and defense are the most important factors in determining a team's points. Further, performance of individual forwards impacts points, but does so inconsistently. More important than individual performance, the number of forwards that a team plays positively impacts points. Finally, goalkeeper performance does not impact team success over the course of a season with statistical significance.

In our sample, MLS teams generally pay strikers according to merit. Strikers that produce and perform well are paid higher salaries. However, teams pay goalkeepers erratically. Considering that goalkeeping has little effect on long-term success, teams would be wise to spend less on a goalkeeper so that they can maximize their budget on other players that influence long-term performance more.

References

American Soccer Analysis. (2018). *Keeper xGoals* [Data file]. Retrieved from https://www.americansocceranalysis.com/keeper-xg-2018/

American Soccer Analysis. (2018). *MLS Player Salaries* [Data file]. Retrieved from https://www.americansocceranalysis.com/may-1-2018/

American Soccer Analysis. (2018). *Player xGoals* [Data file]. Retrieved from https://www.americansocceranalysis.com/player-xg-2018/

BenBaer89. (2018, November 29). #SKCvsPOR xG. Tough one for SKC [Tweet]. Retrieved from https://twitter.com/BenBaer89/status/1068379531445178368

Bogert, T. (2018, December 5). *Atlanta United Ranks Among Global Elite in Average Match Attendance.* Retrieved from https://www.mlssoccer.com/post/2018/12/05/atlanta-united-ranks-among-global-elite-average-match-attendance

*What are Expected Goals?.* American Soccer Analysis. Retrieved from https://www.americansocceranalysis.com/explanation/

Table 2

| | | | | | |
|---|---|---|---|---|---|
| ***Table 2*** | | | | | |
| *Season Performance I* | | | | | |
| Dependent Variable: Total Points in a Season | | | | | |
| *Regressor* | *Model 2.1* | *Model 2.2* | *Model 2.3* | *Model 2.4* | *Model 2.5* |
| Constant | 18.0412 * (6.93) | 47.6647 ** (1.12) | 81.7168 ** (5.60) | 47.4089 ** (1.15) | 50.0426 ** (9.05) |
| xGF (Team Offense) | 0.6468 ** (0.15) | | | | 0.5934 ** (0.13) |
| GSA-xGSA (Primary Forward Performance) | | 0.1841 (0.28) | | | 0.4437 * (0.19) |
| xGA (Team Defense) | | | -0.7551 ** (0.12) | | -0.6697 ** (0.12) |
| GA-xGA (Goalkeeper Performance) | | | | -0.4120 (0.25) | -0.2327 (0.24) |
| SER | 8.446 | 9.473 | 7.900 | 9.335 | 6.909 |
| Adj R^2 | 0.197 | -0.010 | 0.298 | 0.019 | 0.463 |
| Sample includes MLS teams from 2015-2018 with goalkeepers that played at least half a season and had at least 1 forward who played half a season. Standard errors are given in parentheses under coefficients. Coefficients are statistically significant at the *5% or **1% level. | | | | | |

Table 3

| | | |
|---|---|---|
| ***Table 3*** | | |
| *Season Performance II* | | |
| Dependent Variable: Total Points in a Season | | |
| *Regressor* | *Model 3.1* | *Model 3.2* |
| Constant | 46.8428 ** (12.62) | 47.6410 ** (12.18) |
| xGF (Team Offense) | 0.8006 ** (0.16) | 0.7243 ** (0.16) |
| GSA-xGSA (Primary Forward Performance) | 0.3323 (0.28) | 0.3263 (0.26) |
| GSA-xGSA (Secondary Forward Performance) | | 0.8245 ** (0.29) |
| xGA (Team Defense) | -0.8065 ** (0.19) | -0.7774 ** (0.19) |
| GA-xGA (Goalkeeper Performance) | -0.3727 (0.23) | -0.4681 (0.23) |
| SER | 6.568 | 6.197 |
| Adj R^2 | 0.580 | 0.626 |
| Sample includes MLS teams from 2015-2018 with goalkeepers that played at least half a season and had at least 2 forwards who played half a season. Standard errors are given in parentheses under coefficients. Coefficients are statistically significant at the *5% or **1% level. | | |

Table 4

| | |
|---|---|
| ***Table 4*** | |
| *Number of Forwards* | |
| **<u>Dependent Variable: Total Points in a Season</u>** | |
| *Regressor* | *Model 4.1* |
| Constant | 42.5644 ** (1.93) |
| Number of Forwards | 2.6644 * (1.02) |
| SER | 9.487 |
| Adj R^2 | 0.053 |
| Sample includes MLS teams from 2015-2018. Number of Forwards is the number of forwards on a team who played more than half of the minutes in the season Standard errors are given in parentheses under coefficients. Coefficients are statistically significant at the *5% or **1% level. | |

Table 5

| | |
|---|---|
| ***Table 5*** | |
| *Striker Salary* | |
| **Dependent Variable: Salary (USD)** | |
| *Regressor* | *Model 5.1* |
| Constant | -583,630<br>(317,088) |
| GSA - xGSA<br>(Performance) | 91,320 *<br>(43,789) |
| xGSA<br>(Production) | 115,020 **<br>(28,595) |
| SER | 1,299,831 |
| Adj R^2 | 0.210 |
| Sample includes MLS forwards from 2015-2018. Only forwards who played at least half of season are in sample. Performance and production are evaluated per season, not in aggregate. Standard errors are given in parentheses under coefficients. Coefficients are statistically significant at the *5% or **1% level. | |

Table 6

| | |
|---|---|
| ***Table 6*** | |
| *Goalkeeper Salary* | |
| **Dependent Variable: Salary (USD)** | |
| *Regressor* | *Model 6.1* |
| Constant | 262,076<br>(270,590) |
| GA - xGA<br>(Performance) | -7,740<br>(4,878) |
| xGSA<br>(Production) | -119.22<br>(6,617) |
| SER | 371,821 |
| Adj R^2 | -0.018 |
| Sample includes MLS goalkeepers from 2015-2018. Only goalkeepers who played at least half of season are in sample. Performance and production are evaluated per season, not in aggregate. Standard errors are given in parentheses under coefficients. Coefficients are statistically significant at the *5% or **1% level. | |