Capstone Midterm Report: Soccer

Chris Halpert (cph2133), Molly Hanson (meh2243), Rohan Pitre(rp2816), Feng Ye (fy2163)

1 Introduction

In recent years, statistical and analytic methods have been applied within the sports domain, with various parties reaping the benefits:

- Coaches and scouts employ sports analytics to facilitate decision making
- Players can more equally compare their skills and uniquely assess areas of improvement
- Team management can highlight optimal marketing and fan acquisition strategies
- Sports gamblers are gaining from analytics with new information and metrics to assist their betting decisions

Recent sports analytics research primarily uses spatiotemporal data among sports that are easily discretized. Such sports, like tennis, baseball and American football, are simply "broken up into individual events that have immediate outcomes" [1]. Baseball is perhaps the most well-known application of advanced statistics in sports, with Michael Lewis' 2004 novel *Moneyball* that chronicles Oakland A's general manager Billy Beane's use of unique statistics and analytics to build a (winning) team of undervalued players [4].

In sports like basketball and soccer, where "the continuous nature of play makes the connections to outcomes less obvious"[1], comparable analytics work is more challenging. Nonetheless, basketball is evolving with the new "Player Tracking" technology [5], which is improving team efficiently through player movement analysis. Soccer, which has even weaker "distinctions between epochs in play" compared to basketball [1], is one of the last sports to fully embrace analytics.

By applying analytics to soccer, there is the ability to improve the level of game play and challenge existing myths [2]. Research in behavioral economics explains us the tricks our minds play on us, and in turn how this affects our decision making. Data can be used to annul various myths that are based off cognitive biases, and provide grounds for rational decisions. For instance, [2] discounts the notion of increased vulnerability of a team after scoring and proves that the number of goals a team scores is not positively related to the number of corners it wins. It is important to utilize data for decisioning, as many beliefs about the game are not necessarily proven. Moreover, as seen with the Moneyball approach, analytics can be used to highlight undervalued soccer players.

2 Project Goals

The goal of our capstone project is to predict the performance and prices of soccer player. This project is unique in that betting algorithms largely consider team performance, whereas the scope of our project will focus on player level predictions. This model we produce is important as teams are interested in their return on investment: contracting cheap players with the best possible performance. Thus, it is imperative that our model be highly interpretable, whereby the inputs and output are common metrics or variables that teams can measure on their own.

It is also important to note that the idea of *good* performance will differ depending on a player's position. Therefore, our model must consider this variable and the output must correspond to the metric based on the given player's position. Lastly, our project should highlight which variables are less correlated with salary but correlated with performance, in order to suggest highly skilled, yet undervalued players.

3 The Data

3.1 Overview

3.2 Challenges

4 Initial Data Exploration

5 Related Literature

5.1 Performance

Soccer analysts are using several new, advanced metrics to predict team and player performance. Caley [6] introduced expected goals, xG, to "estimate the quality of chances that a soccer team creates and concedes in a match". Caley identifies six types of shots, including regular shots, headed shots not assisted by crosses and shots from direct free kicks, and employs logistic regression to produce a distinct xG model for each shot type. Input variables include angle and distance to the net, how the shot was assisted, type of attacking play, and various other indicators for player finishing skills and defensive pressures. While xG is a highly intuitive predictive model, [7] argues that Caley's model is poorly constructed, and the r-squared of 0.997 is highly suspicious and would be a poor metric to assess the explanatory capabilities of the model in non-linear instances.

Likewise, Eastwood [8] trained an xG model using support vector machines, which can be used with non-linear data, and to account for the variability of the metric, he used a bootstrapping technique to provide a confidence interval for range of values the expected goals is likely to fall within. Further, [8] measures the error of his model using RMSE, however RMSE must be baselined and compared to similar models to truly have meaning.

Expected points [9] takes the xG model one step further to account for the inequality of all goals, with the aim to measure the importance of a goal is and can highlight players who take more important shots, in turn impacting team performance. Notably, while xG and expected points may be useful in predicting the performance of a forward, it would not be fair to use this metric to assess the performance of a defenseman since their primary role does not involve goal scoring. Since our model goal includes predicting individual player performance, xG may be considered a performance metric for forwards, however other statistics must be used for players in other positions.

With the exception of the goalkeeper, successful passes is a common performance metric for all positions. Using locations of pass data, [1] present a "novel way of quantitatively measuring a player's passing performance", under the assumption that "pass location is a strong indicator of team strategy and personnel". This model first discretizes each game in the data set into a discrete event sequence by segmenting games into possessions. An L2-regularized SVM approach was used to model the data and outputs player rankings by shot predictions using Pass Shot Value. Notably, computing each players Average Pass Shot Value can be used to compare players within each position category (offense, midfield, defense) for a more valid comparison.

Lastly, it is vital to remember that luck plays a large role in soccer, especially compared to other sports. Reep [10] observes the stochastic element in the "number of goals arising from a particular number of shots in one match", indicating that simply because one team shoots more, it does not mean the other team will not get more goals. This observation that "the odds of a pass being completed at any one time are no more than 50/50" [2] demonstrates how much soccer relies on both skill and fortune equally. Teams must account for this randomness in coaching, and likewise, analysts must not overlook this reality in modeling practices.

5.2 Pricing

Research into the pricing models of soccer players seeks to identify variables that are strong predictors of salary. Studies have shown [12] an effect of salary on performance, but no effect of performance. Additionally, some of the observable variations in player salaries can be explained by two-footedness, however [3] finds the European leagues to efficiently price two-footedness. Broader studies consider linear modeling techniques to predict the market value of soccer players, namely [11] considers 16 predictors among three categories:

- Personal player information, including position, height, foot
- Performance data, including division, goals, appearances
- Ratios of predictors, including goal rate and an indicator of when a player became famous in his career

Using different regression modeling techniques including OLS, KNN, Ridge Regression, Principal Component Regression, along with 10-fold validation, [11] finds the most important variables in predicting player market value to be: position, team rank,

foot, height, and age. This model, however, neglects the notion that different positions should be held to different performance metrics.

Additionally, [13] also considers the *superstar effect* and segments player salaries into percentiles to see the differing effects of variables on pricing at various percentiles. For instance, [13] concludes that only by the 98% quantile does each incremental goal increase the transfer value, whereas assists are statistically by the 90% quantile. Other related research [12] observes a link between lower overall team performance and wider salary spreads on a team, which is important for management to consider in their pricing strategies.

6 Proposed Solution

References

[1] J. Brooks, M. Kerr, and J. Guttag. Developing a data-driven player ranking in soccer using predictive model weights. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* 2016.

[2] C. Anderson and D. Sally. *The numbers game: why everything you know about soccer is wrong.* New York: Penguin Books, 2014.

[3] A. Bryson, B. Frick, and R. Simmons. The returns to scarce talent: footedness and player renumeration in European soccer. *Journal of Sports Economics,* 14(6):606-628, 2012.

[4] M. Lewis. *Moneyball: the art of winning an unfair game.* New York: W. W. Norton & Company, 2004.

[5] L. Steinberg. Changing the game: the rise of sports analytics. *Forbes,* August 2015.

[6] M. Caley. Premier league projects and new expected goals. *SBNation,* October 2015.

[7] M. Bertin. Why soccer's most popular advanced stat kind of sucks. *Deadspin,* April 2015.

[8] M. Eastwood. Expected goals and support vector machines. *pena.lt/y,* July 2015.

[9] J. Young. Goodbye expected goals, hello expected points! *American Soccer Analysis,* July 2016.

[10] C. Reep and B. Benjamin. Skill and chance in association football. *Journal of the Royal Statistical Society. Series A (General),* 131(4):581-585, 1968.

[11] Y. He. Predicting market value of soccer players using linear modeling techniques. *Berkely University,* 2014.

[12] B. Torgler and S. Schmidt. Relative income position and performance: An empirical panel analysis. *Applied Economics,* 37, 2355-2369, 2007.

[13] D. Newman. Predicting transfer values in the english premier league. *Duke University,* 2016.