MAYFIELD: Machine Learning Algorithm for Yearly Forecasting Indicators and Estimation of Long-Run Player Development

Accurate statistical prediction of American football player development and performance is an important issue in the sports industry. We propose and implement a novel, fast, approximate k-nearest neighbor regression model utilizing locality-sensitive hashing in highly dimensional spaces for prediction of yearly National Football League player statistics. MAYFIELD accepts quantitative and qualitative input data, and can be calibrated according to a variety of parameters. Concurrently, we propose several new computational metrics for empirical player comparison and evaluation in American football, including a weighted inverse-distance similarity score, stadium and league factors, and NCAA-NFL statistical translations. We utilize a training set of comprehensive NFL statistics from 1970-2019, across all player positions and conduct cross-validation on the model with the subset of 2010-19 NFL statistics. Preliminary results indicate the model to significantly improve on current, publicly available predictive methods. Future training with advanced statistical datasets and integration with scouting-based methods could improve MAYFIELD's accuracy even further.

1 Introduction

Accurate forecasting of on-field performance in professional sports is a major component of player evaluation by fans, coaches, and sports executives. Due to a variety of factors, including a lack of highly detailed, publicly available data, emphasis on traditional video scouting methods by coaches and executives, and the relative complexity of the sport, American football is considerably less analytically developed than other professional sports, most notably baseball and basketball. Silver (2003, 2015) presents advanced comprehensive forecasting models for the MLB and NBA, although these algorithms are proprietary and are thus irreproducible. Schatz (2008) presents a similar non-comprehensive[1] model for the NFL, although it is likewise not publicly available.

We present a reproducible, comprehensive, learning-based methodology for year-by-year statistical forecasting of NFL players' careers, and implement it on the entire set of post-merger (i.e., after 1970) NFL players. A wide survey of the relevant literature reveals that, to date, no algorithm exists which comprehensively projects NFL player statistics across all positions, and utilizes a dataset of MAYFIELD's size and scope. We also propose several important contributions to football analytics for future implementation into MAYFIELD: an Approximate Value metric for collegiate football players, NCAA-NFL statistical translations which adjust for park and league factors, and a Jamesean-style Similarity Scores framework for empirical player comparison.

2 Methodology

2.1 Data

MAYFIELD utilizes a dataset derived from the Pro Football Reference (PFR) historical database consisting of on-field performance statistics and biographical information for every player, team, and season since 1970, when the NFL and the AFL merged. For purpose of analysis, we segregate the player data into several position groups according to their PFR-listed positions as shown in Figure 1[2].

[1] Offensive linemen and punters are not included in Schatz's (2008) forecasts.

[2] Note that some PFR-listed positions are mapped to MAYFIELD's according to the defensive schemes which the player performed under (i.e., 4-3 or 3-4). This is denoted by the name of the scheme marked on the corresponding arrow.

Our on-field statistics comprise a set of standard NFL box-score statistics such as yards, touchdowns, tackles, field goal attempts, etc. As depicted in Figure 2, we limit the set of variables considered for players at a given position to only those relevant to the on-field role of a typical player at that position. Importantly, we include PFR's "Approximate Value" metric (Drinen, 2008a), which places a numerical value on the all-inclusive contributions of a player to his team's success in a given season.

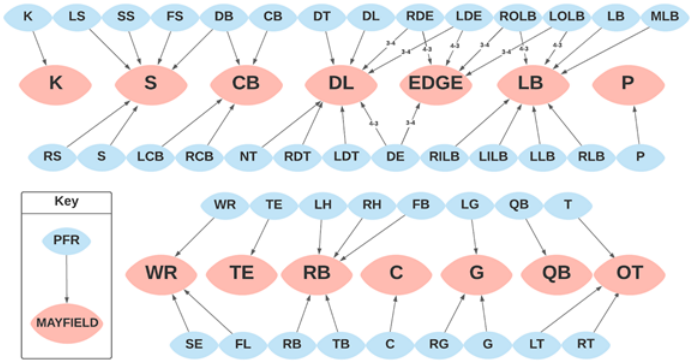

Figure 1: Position Mapping

Our biographical data can be split into two categories: static, and dynamic. Static biographical variables are variables for a player whose initial value never changes from year to year, whereas dynamic biographical variables may fluctuate from year to year. We list which biographical variables are used for offensive (i.e., QB, RB, TE, WR, OL) and defensive (i.e., DL, EDGE, LB, CB, S) players in Figure 3[3] In addition to data collected from PFR, some of the dynamic variables are not taken explicitly from PFR, but instead calculated from the data, including the variables for changes coaching, team, and scheme, consecutive years with a players' current coaches, team, and scheme, and the total AV per game of a team's players at each position during the previous season. One dynamic biographical statistic of note is team ratings from the Simple Rating System (SRS), which is described by Drinen (2006). SRS estimates the strength of a team's offense and defense relative to the league average in terms of points per game- for instance, a team with an offensive SRS of +6.0 would be expected to score 6 more points per game than an average team, all else equal.

Figure 2: Performance Variable Assignment

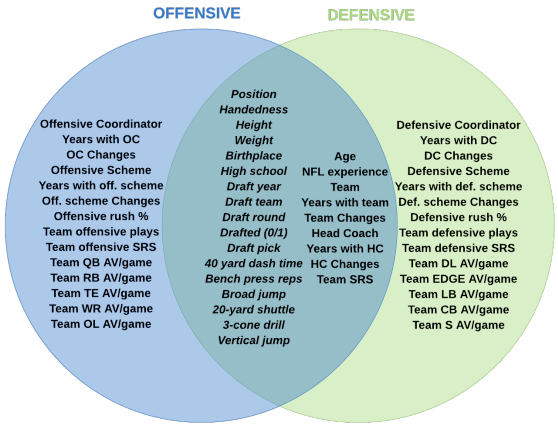

Figure 3: Biographical Variable Assignment

[3]Special teams players (K, P) only use the statistics listed in the intersection of Figure 3. Note that static biographical statistics are italicized.

2.2 The kNN Algorithm

K-nearest neighbor (kNN) algorithms are among the oldest and most widely utilized methods in machine learning, and have been shown as versatile, but surprisingly effective methods for classification and regression problems. Importantly, several advanced player projection techniques similar to MAYFIELD (Silver, 2003, 2015; Schatz, 2008) utilize nearest-neighbor comparisons, although their proprietary nature prevents analysis of their algorithmic structure (kNN or otherwise). MAYFIELD utilizes a similar technique as these algorithms, albeit with more formalized and reproducible methods.

MAYFIELD generates player statistic predictions in the following manner. Given the entire training set of Y-length segments S and values for each hyperparameter, we specify a subset $R = \{r^{i,t} \in S : \exists r^{i,t+u} \in S\}$ of segments which have corresponding future performance vectors in the training data. We then conduct direct nearest-neighbor search on R from which the set $H_{r^{i,t}}$ of K most similar segments to each $r^{i,t}$ are then used to predict via LOESS regression the future performance vectors $\{v^{i,t+u}\}$ which correspond to R. We then conduct cross-validation and select the most optimal model across each different combination of hyperparameters, a procedure we describe in Section 3.1.

2.2.1 Inverse Euclidean Distance Similarity Scores

The principal question when designing a nearest-neighbor search algorithm is that of similarity measurement. The standard approach (which we adopt) is to specify some distance metric on the search set, and to identify the objects with lower distances to the object of interest as monotonically more similar.

Similarity measures appear in some sabermetrics and other sports analytics work, albeit in a more informal manner than the academic literature. Bill James (1994) first introduced the concept of "similarity scores" for empirical comparison of Baseball Hall of Fame candidates to its already inducted members, which spawned a variety of methods for sports performance comparison (e.g. Silver, 2015; Kubatko, 2004; Hollinger, 2003; Pelton, 2003; Drinen, 2008b). Some forecasting models (Silver, 2003, 2015; Schatz, 2008) incorporate similarity scores into their respective methodologies as well. Similarity scores are one of the few areas of advanced empirical analysis of sports in which American football is well-represented. Schatz (2010) and Drinen's (2008b) methodologies are well-known, and more importantly, reproducible. However, almost all work on sports performance similarity scores have used strictly first or second order polynomial models[4] with no consideration for side information such as that of our biographical variables, which describe potentially predictive factors at the player and team levels. We seek to rectify these shortcomings by returning to more formally established methods in the non-sporting literature.

Computing distances between segments requires two separate metrics: one for quantitative data, and the other for categorical values. We select the standardized Euclidean norm and the composite Jaccard distance (Jaccard, 1901) for these, respectively. So, where we have an n-dimensional quantitative feature space, and an m-dimensional categorical feature space, we define our distance function d over the whole space as:

$$d(r^{i_1,t}, r^{i_2,t}) = \sqrt{\sum_{j=1}^{n} \frac{(r_j^{i_1,t} - r_j^{i_2,t})^2}{\sigma_j^2}} + \sum_{j=n+1}^{n+m} (1 - \frac{|r_j^{i_1,t} \cap r_j^{i_2,t}|}{|r_j^{i_1,t} \cup r_j^{i_2,t}|}) \tag{1}$$

d is just the sum of the standardized Euclidean metric over the quantitative features of $r^{i_1,t}$ and $r^{i_2,t}$ and the Jaccard distance between the categorical features of $r^{i_1,t}$ and $r^{i_2,t}$. We then define the similarity

[4]The methods we review exclusively use affine equations over either absolute or squared differences in observed feature values to compute their similarity scores.

measure between $r^{i_1,t}$ and $r^{i_2,t}$ as the radial-basis function (RBF) kernel of d:

$$SS(r^{i_1,t}, r^{i_2,t}) = e^{-\frac{1}{2}(\frac{d(r^{i_1,t}, r^{i_2,t})}{n+m})^2} \tag{2}$$

The RBF kernel is a natural choice of similarity measure due to our utilization of the Euclidean metric as a component of d and since the average distance $\overline{d}$ will always be close to the dimension of the feature space, $n + m$. This parameterization is both easily interpretable and flexible enough to admit insertion of additional variables into the feature set without the need to re-specify a new distance or similarity function, and is easily modified to weight the various features (simply multiply the jth terms in each summation by some β_js) , although we do not do so at this time. For every segment $r^{i,t} \in R$, we compute the following set of $r^{i,t}$'s "nearest neighbors":

$$H_{r^{i,t}} = \{r^{i_1,t} \in R : |\{r^{i_2,t} \in R : SS(r^{i_2,t}, r^0) \geq SS(r^{i_1,t}, r^0) \wedge i \neq i_2\}| \leq K \wedge i \neq i_1\} \tag{3}$$

Less formally, $H_{r^{i,t}}$ is just the set of the K most similar segments to $r^{i,t}$, excepting those segments which describe the same player as $r^{i,t}$. We now proceed to the following steps in the MAYFIELD algorithm, where we describe MAYFIELD's further utilization of $H_{r^{i,t}}$ in its predictions of $v^{i,t+u}$.

2.2.2 The Regression Equation

Given the set $H_{r^{i,t}}$, our remaining task is to form an estimate of $v^{i,t+u}$. Consider the set $G_{r^{i,t}}$, the set of future performance vectors which correspond to the members of $H_{r^{i,t}}$:

$$G_{r^{i,t}} = \{v^{i,t+u} \in V : r^{i,t} \in H_{r^{i,t}}\} \tag{4}$$

Classical kNN regression techniques (e.g. Benedetti, 1977; Altman, 1992) typically compute a weighted average of vectors in $G_{r^{i,t}}$ as the predicted value of $v^{i,t+u}$, an approach we adopt in our formulation of $v^{i,\hat{t}+u}$:

$$v^{i_0,\hat{t}+u} = \frac{\sum_{v^{i,t+u} \in G_{r^{i_0,t}}} SS(r^{i,t}, r^{i_0,t})^2 v^{i,t+u}}{\sum_{v^{i,t+u} \in G_{r^{i_0,t}}} SS(r^{i,t}, r^{i_0,t})^2} \tag{5}$$

$v^{i_0,\hat{t}+u}$ is just an inverse-distance weighted interpolation of G_{r^0}. Note that any features for which either segment has missing data are dropped from the summations.

However, we are not quite done. Dependent on the local structure of the data, systematic bias in $v^{i_0,t+u}$ may arise among any or all of the components of V. For instance, $v_j^{i_0,t+u}$ may be nonmonotonic with respect to $v_j^{i_0,\hat{t}+u}$, or may increase at a rate different from unity. kNN regressions such as MAYFIELD are particularly suspect in this respect, since interpolations on the data as performed in nearest-neighbor algorithms may over-predict regression to the mean. Although we expect such cases to be limited, robustness to such issues is a desirable feature for regression models.

Given the possible estimation bias that may enter into our regression, we take the additional step of fitting a LOESS model for each feature in V, i.e., regressing $v_j^{i,t+u}$ on $v_j^{i,\hat{t}+u}$. LOESS (Savitzky & Golay, 1964; Cleveland, 1979), or locally estimated scatterplot smoothing, is a nonparametric method which estimates weighted low-order polynomial regressions on overlapping subsets of the independent variable, and constructs a smoothed function on the local polynomials. The size of these subsets depends on the bandwidth hyperparameter B, which is the ratio of the order of the subsets to N, the number of total players in

R. Larger values of B result in smoother results, and make the locally estimated polynomials more robust to outliers. We set $B = \frac{log_2(N)}{N}$, which results in a relatively large bandwidth[5], as our data is both highly noisy and highly dense. So, where g_j is the LOESS-constructed function for the jth feature of V, our final predicted value of the performance vector $v^{i,t+u}$ is:

$$\hat{v^{i,\hat{t}+u}} = \sum_{j=1}^{p} g_j(v_j^{i,\hat{t}+u})\hat{e_j} \tag{6}$$

Our utilization of this LOESS correction to $\hat{v}^{i,t+u}$ yields several advantages. Firstly, the aforementioned issues of bias in $v^{i,\hat{t}+u}$ are resolved. Secondly, the nonparametric nature of LOESS admits a far more flexible bias correction than the standard parametric models; cases of non-monotonically increasing $v^{i,t+u}$ with respect to $v^{i,\hat{t}+u}$ are in particular likely to benefit. Thirdly, LOESS's confidence intervals are calculated based on the local structure of the data, as opposed to aggregate measures of variance (as in most least-squares regressions), and give a meaningful and ready-made confidence bounds on the expected range of $v^{i,t+u}$. These may be especially useful when more than just point-estimates of $v^{i,t+u}$ are required. Finally, investigation on the shape of the LOESS-generated functions g_j may reveal various characteristics of $v^{i,\hat{t}+u}$ such as over or under-prediction of regression to the mean, artificially induced clustering, or lack of predictive power by the training data.

3 Results

4 Future Work

The MAYFIELD algorithm we propose is designed in a manner which supports integration with possible future advances in football analytics. For instance, while our dataset consists of only standard box-score performance variables, newer metrics reliant[6] upon player-tracking or other advanced techniques are easily absorbed into MAYFIELD's feature space, since our distance function is robust to missing data (a major concern for any newer metrics which are unable to be calculated for historical data). Similarly, integration with other contextual data, such as player positions on depth chart, scouting report grades, and medical or salary information, would likewise increase MAYFIELD's accuracy by their addition to the feature space without any modification needed to make full use of the inserted variables.

Technical modifications to augment MAYFIELD's predictive abilities are likewise possible without departing from the framework we present. Insertion of learned feature weights to the distance function, adjustment for feature covariance via the generalization of the quantitative feature space metric to a Mahalanobis distance, increased identification of terms in the first-stage regression equation, and expanded testing of optimal values for Y and K all could potentially improve MAYFIELD's already-impressive accuracy, compared to other currently existing models.

One important component of MAYFIELD's evolving methodology will be the inclusion of player data from amateur, collegiate, and other professional football leagues. We propose a methodology for stadium, league, and year factors for future use in MAYFIELD or other algorithms based off of the work of Szymborski (1997); Davenport (1996); James (1985), and Thorn & Palmer (1984). To calculate stadium factors of each

[5]In some specialized cases where there are abnormally high frequencies of zeros in the data, we increase B to avoid failure of local polynomial estimations due to lack of variance in $v_j^{i,\hat{t}+u}$.

[6]E.g., pass blocking and pass rushing win rate (Burke, 2018), air yards, completion percentage allowed, nflWAR (Yurko, Ventura & Horowitz, 2019), etc.

statistic for each team and season , we propose a modified variant of Thorn & Palmer (1984)'s calculations. Our base factor Φ for team i in year t on a statistic λ where $\lambda^{i,t}$ is the per-game average of λ which team i recorded in year t, and $\lambda'^{i,t}$ is the per-game average of λ which team i allowed in year t, is:

$$\Phi^{i,t} = \frac{\lambda^{i,t}_{home} + \lambda'^{i,t}_{home}}{\lambda^{i,t}_{away} + \lambda'^{i,t}_{away}} \tag{7}$$

However, Φ is not entirely satisfactory, since it accounts for neither the caliber of the team in question nor its opponents. To resolve the issue, we introduce Thorn & Palmer's (1984) offensive and defensive team ratings with respect to λ which we call τ. The formulae for τ are as follows, letting n^t be the number of NFL teams playing during the current year:

$$\tau^{i,t}_{off} = [\frac{\lambda^{i,t}_{away}(n^t-1)}{n^{2t}-2n^t+\Phi^{i,t}} + \frac{\lambda^{i,t}_{home}}{\Phi^{i,t}(n^t-\Phi^{i,t})}] \cdot \frac{n^t - 2 + \tau^{i,t}_{def}}{2\overline{\lambda^{i,t}}} \;\; ; \;\; \tau^{i,t}_{def} = [\frac{\lambda'^{i,t}_{away}(n^t-1)}{n^{2t}-2n^t+\Phi^{i,t}} + \frac{\lambda'^{i,t}_{home}}{\Phi^{i,t}(n^t-\Phi^{i,t})}] \cdot \frac{n^t - 2 + \tau^{i,t}_{off}}{2\overline{\lambda^{i,t}}} \tag{8}$$

Since τ are codependent, we initialize each $\tau = 1$, and then recompute their values until convergence, typically after 3 iterations. The adjusted stadium factors Ω are then respectively:

$$\Omega^{i,t}_{off} = \frac{n^t - \Phi^{i,t}\frac{2n^t-\Phi^{i,t}-1}{2n^t-2}}{n^t - 2 + \tau^{i,t}_{def}} \;\; ; \;\; \Omega^{i,t}_{def} = \frac{n^t - \Phi^{i,t}\frac{2n^t-\Phi^{i,t}-1}{2n^t-2}}{n^t - 2 + \tau^{i,t}_{off}} \tag{9}$$

We maintain Thorn & Palmer's (1984) segregation of offensive and defensive factors in our calculations; that is, offensive players' statistics are adjusted using Ω_{off} and defensive players' statistics are adjusted using Ω_{def} However, instead of the typical 1-year park factors used in sabermetrics, we suggest an unweighted 5-year moving average of Ω due to the comparably smaller sample size of games in a given NFL season. Without adjustment to the time horizon of stadium factor calculations, random variance in the data may influence stadium factors more than is optimal.

So far, our stadium factors only differ from those of Thorn & Palmer (1984) in our calculation of Φ From here, we utilize Szymborski's (1997) method for our translations. The two remaining components of interest are year and league factors. To adjust for these effects, we compute a league-year factor, Θ for λ as follows:

$$\Theta^t = \frac{\overline{\lambda^{i,*}}}{\overline{\lambda^{i,t}}} \cdot \delta \tag{10}$$

Θ is just the average team's λ during some base year and league over the average team's λ in a given league and year, times a deflator δ which measures the relative talent level of a player's league as a compared to the base league. The calculation of δ is not detailed by Szymborski, so where $\chi^{i,t}$ is the cumulative yearly statistics which player i recorded, we define δ as the ratio in NFL-average $\chi^{i,t}$ and the average $\chi^{i,t}$ for a given league after adjustment for Ω and Θ, among players which played in both the NFL and that league over the course of their careers. The final translated value of a player's $\chi^{i,t}$ is then:

$$\hat{\chi^{i,t}} = \chi^{i,t}\sqrt{\frac{\Theta^t}{\Omega^{i,t}}} \tag{11}$$

These translations are not a prediction of how a player would have performed during that year if they had been in the NFL, rather, they are an estimation of the player's observed performance in the context of the baseline NFL environment. Changes in playing time due to higher levels of team talent, differences in play style, coaching, and strategy of NFL teams as opposed to collegiate teams, and other factors which

influence observed player statistics in more nuanced manners are not captured by these translations.

5 Conclusion

References

Altman, N.S. 1992. An introduction to kernel and nearest neighbor nonparametric regression. The American Statistician, 4(3): 175-185.

Benedetti, J. K. 1977. On the Nonparametric Estimation of Regression Functions. Journal of the Royal Statistical Society Series B, 39(2): 248-253.

Blees, C. 2011. Running Backs in the NFL Draft and NFL Combine: Can Performance be Predicted? Claremont McKenna College. Accessed at https://scholarship.claremont.edu/cgi/viewcontent.cgi?article=1180&context=cmc_theses.

Burke, B. 2018. We created better pass-rusher and pass-blocker stats: How they work. ESPN. Accessed at https://www.espn.com/nfl/story/_/id/24892208/creating-better-nfl-pass-blocking-pass-rushing-stats-analytics-explainer-faq-how-work.

Cleveland, W.S. 1979. Robust locally weighted regression and smoothing scatterplots. Journal of the American Statistical Association, 74(368): 829-836.

Davenport, C. Davenport Translations; in Huckaby, G., Davenport, C., Jazayerli, R., Kahrl, C., Sheehan, J. 1996. Baseball Prospectus 1996. Baseball Prospectus LLC. Accessed at https://legacy.baseballprospectus.com/other/bp1996/dtessay.html.

Drinen, D. 2006. A very simple ranking system. Pro Football Reference. Accessed at https://www.pro-football-reference.com/blog/index4837.html?p=37.

Drinen, D. 2008. Approximate value in the NFL. Pro Football Reference. Accessed at https://www.pro-football-reference.com/blog/index6b92.html?p=465.

Drinen, D. 2008. Who is the current Dave Duerson?. Pro Football Reference. Accessed at https://www.pro-football-reference.com/blog/indexa215.html?p=556.

Hollinger, J. 2003. Pro Basketball Prospectus: 2003 Edition. University of Nebraska Press. Print.

Jaccard, P. 1901. Etude comparative de la distribution florale dans une portion des Alpes et du Jura., Bulletin de la Societe Vaudoise des Sciences Naturelles, 37(1): 547–579.

James, B. 1985. The Bill James baseball abstract, 1985. Ballantine Books. Print.

James, B. 1994. The politics of glory. Macmillan Publishers. Print.

Kapania, N. 2012. Predicting Fantasy Football Performance with Machine Learning Techniques. Stanford University. Accessed at http://cs229.stanford.edu/proj2012/Kapania-FantasyFootballAndMachineLearning.pdf.

Kubatko, J. 2004. Similarity scores. Basketball Reference. Accessed at https://www.basketball-reference.com/about/similar.html.

Mulholland, J., Jensen, S. 2018. Predicting the Future of Free Agent Receivers and Tight Ends in the NFL. Statistica Applicata - Italian Journal of Applied Statistics Vol. 30 (2). Accessed at http://www-stat.wharton.upenn.edu/ stjensen/papers/shanejensen.football.freeagents.2018.pdf.

Pelton, K. 2003. Review: Pro Basketball Prospectus: 2003-04 Edition. Hoopsworld. Accessed at http://www.hoopsworld.com/article_5978.shtml.

Porter, J. 2018. Predictive Analytics for Fantasy Football: Predicting Player Performance Across the NFL. University of New Hampshire. Accessed at https://scholars.unh.edu/cgi/viewcontent.cgi?article=1418&context=honors.

Savitzky, A., Golay, M.J.E. 1964. Smoothing and differentiation of data by simplified least squares procedures. Analytical Chemistry, 36(8): 1627–1639.

Schatz, A. 2008. Pro Football Prospectus 2008. Plume. Print.

Schatz, A. 2010. Football Outsiders similarity scores. Football Outsiders. Accessed at https://www.footballoutsiders.com/stats/similarity.

Silver, N. Introducing PECOTA; 507-514 in Huckaby, G., Kahrl, C., Pease, D. 2003. Baseball Prospectus: 2003 Edition. Potomac Books. Print.

Silver, N. 2015. We're predicting the career of every NBA player. Here's how. FiveThirtyEight. Accessed at https://fivethirtyeight.com/features/how-were-predicting-nba-player-career/.

Szymborski, D. 1997. How to calculate MLEs. Baseball Think Factory. Accessed at https://www.baseballthinkfactory.org/btf/scholars/czerny/articles/calculatingMLEs.htm.

Thorn, J., Palmer, P. 1984. The hidden game of baseball. Knopf Doubleday Publishing Group. Print.

Yurko, R., Ventura, S., Horowitz, M. nflWAR: a reproducible method for offensive player evaluation in football. Journal of Quantitative Analysis in Sports, 15(3): 163-183.