10-25-2019

Improving Foresight Predictions in the 2002-2018 NFL Regular-Seasons: A Classic Tale of Quantity vs. Quality

Eduardo C. Balreira
Trinity University, ebalreir@trinity.edu

Brian K. Miceli
Trinity University, bmiceli@trinity.edu

Journal of Advances in Mathematics and Computer Science

34(1): 1-14, 2019; Article no.JAMCS.52105

ISSN: 2456-9968

(Past name: British Journal of Mathematics & Computer Science, Past ISSN: 2231-0851)

Improving Foresight Predictions in the 2002–2018 NFL Regular-Seasons: A Classic Tale of Quantity vs. Quality

E. Cabral Balreira[1*] and Brian K. Miceli[1]

[1] *Department of Mathematics, Trinity University One Trinity Place San Antonio, TX 78212-7200, United States.*

Authors' contributions

This work was carried out in collaboration between both authors. Both authors read and approved the final manuscript.

Article Information

DOI: 10.9734/JAMCS/2019/v34i1-230203

Editor(s):

(1) Dr. Dragos-Patru Covei Professor, Department of Applied Mathematics, The Bucharest University of Economic Studies, Romania.

Reviewers:

(1) Mudasir M Kirmani, Sher-e-Kashmir University of Agricultural Sciences and Technology, India.

(2) Tunjo Perić, University of Zagreb, Croatia.

Complete Peer review History: http://www.sdiarticle4.com/review-history/52105

Received 11 August 2019

Accepted 19 October 2019

Published 25 October 2019

Original Research Article

Abstract

Utilizing a modified Bradley-Terry model, we develop a method of making foresight predictions of 2002–2018 NFL games by incorporating a home-field parameter into previously established ranking models. Knowing only the home team and score of each contest, and taking into account previous predictions, we optimize this parameter considering one of two things: the quantity of correct picks to date or the quality of predictions to date as measured by a quadratic scoring function. Our main results establish that optimization of quality—rather than quantity—when making a prediction has higher overall accuracy.

Keywords: Bradley-Terry; NFL; rankings; foresight predictions; scoring functions.

**Corresponding author: E-mail: ebalreir@trinity.edu;*

1 Introduction

One of the great joys of NFL regular-season fandom is sitting down each week to predict the outcomes of upcoming games. The major television networks have season-long competitions amongst their broadcasters to see who is the best at correctly picking games, and of course, it makes colloquial sense to equate *accurately* picking games with making *quality* picks. However, in a mathematical sense, the assessment of the quality of picks is distinct from the quantity of correct picks. For instance, suppose there are three games to be picked—A vs. B, C vs. D, E vs. F—where the winners were A, C, and E. Now further suppose that before these games are played, two players—Player 1 and Player 2—predict the outcomes of each game and also assign a confidence to each prediction. Consider the example in Table 1 where Player 1 chooses all three outcomes correctly, while Player 2 picks A and C to win, but also incorrectly chooses F to beat E. From a purely quantitative standpoint, we can say Player 1 picks better than Player 2 in this example.

Table 1. Example of game predictions illustrating quantity vs. quality of correct picks.

Player	Game	Pick	Prediction Confidence
Player 1	A vs. B	A	60%
	C vs. D	C	55%
	E vs. F	E	51%
Player 2	A vs. B	A	90%
	C vs. D	C	80%
	E vs. F	F	55%

However, as we consider the probability assigned to their predictions, we see that while both players choose correctly in the first two games, Player 2 is quite confident about the actual outcome of those games; in the one game Player 2 got incorrect, they are not especially certain of the outcome. This is in comparison with Player 1, who picks correctly all three games, yet is not especially confident about the outcome of any particular game. What if now Players 1 and 2 are picking a fourth game, say between teams G and H, where Player 1 chooses G with 60% confidence and Player 2 picks H with 70% confidence? Who do we feel is more likely to have the pick correct? The purpose of this paper is to look at whether or not the foresight pick accuracy of a model can be improved by taking in consideration the quality of previously made predictions instead of just focusing on the right-wrong accuracy of the previous predictions. We will show that by considering home field advantage and focusing on quality of predictions rather than quantity we can actually improve the overall accuracy of our predictions of regular-season NFL games.[1]

The outline of this paper is the following. In Section 2 we briefly describe various rating methods, and we discuss how we use the given rating methods to assign a prediction probability to each game played. We also describe the idea of *pick quality* as being something distinct from *pick accuracy*, and we show how to incorporate the idea of quality of previous picks into making picks for an upcoming game. In Section 3 we present the results of picking regular-season NFL games using various methods, and in Section 4 we discuss the ramifications of using pick quality with respect to improving pick accuracy.

[1]Since the last NFL expansion was in 2002, we limit the scope of the analysis to the 2002–2018 NFL seasons, although the methodology employed is certainly transferrable to other settings and time frames.

2 Rating Systems & Methodology

Suppose we have N teams, denoted as $T_1, T_2, \ldots, T_N$, competing in an m-round tournament, where in any given round each team plays at most one other team. After the k-th round of the tournament has been played, we assign to each T_i a *rating* $r_i^{(k)} \in \mathbb{R}$, and we set $\mathbf{r}^{(k)} = \left(r_1^{(k)}, r_2^{(k)}, \ldots, r_N^{(k)}\right) \in \mathbb{R}^N$ to be the *k-th rating vector*—the idea here is that the value of $r_i^{(k)}$ is some reflection of T_i's quality after the k-th round. We then use $\mathbf{r}^{(k)}$ to formulate a prediction probability of the game outcomes for the k+1-st round of the tournament. Specifically, we wish to assign a probability to the prediction that T_i beats T_j in the k+1-st round of the tournament, and this probability should depend on $r_i^{(k)}$ and $r_j^{(k)}$. Accordingly, two implicit issues arise. First, we must determine the numerical values of the rating vector $\mathbf{r}^{(k)}$, that is, we must decide on a *ranking method* for the tournament. Second, we must decide on a suitable function to assign a probability to the outcome of each game in round k+1, including whether or not we consider a home field advantage.

The literature is full of ranking methods, and for detailed descriptions of the majority of the methods discussed in subsequent sections, we refer the reader to [1]. For our purposes, we separate ranking methods into four categories: probabilistic, win/loss (W/L), deterministic, and Markov. We will denote by *ω-type* methods those that only take into consideration the outcome of the games and *σ-type* methods those that take into account the scores of games played. The former case may be considered as a special case of the latter by treating every game as if the final score is either 1-0, 0-1, or 0.5-0.5, and we use this method to construct ω-type ratings for models that typically incorporate game scores into their ratings.

2.1 Bradley-Terry

The Bradley-Terry model, developed in [2], is one of the more influential methods of producing ratings, and indeed, the probabilistic method we employ as the basis for the rest of this paper is a slightly modified version of Bradley-Terry. Generally speaking, Bradley-Terry estimates the probability that T_i beats T_j to be

$$p_{ij} = \frac{r_i}{r_i + r_j}, \tag{2.1}$$

where r_i and r_j are ratings of T_i and T_j, respectively. The way in which Bradley-Terry assigns these rating is by computing p_{ij} as the actual observed likelihood that T_i beats T_j, and then finding the coordinates of the rating vector $\mathbf{r}$ using maximum likelihood estimates (MLE), see [3]. We implement this method here using a simple iterative algorithm from the work of [4], but to implement Bradley-Terry for the NFL, two remarks are necessary. First, the existence of the MLE is contingent on the strong connectivity of the associated network, as observed in [5]. Related directly to the NFL, this may not always happen as there are seasons with undefeated or winless teams until very late in, or even to the end of, a season. Second, since during an NFL regular season two teams will only play twice, once, or zero times against one another, the observed likelihood of a win can only be 1, 0.5, or 0. A solution introduced in [6] uses the scores of each game to provide a different estimate of p_{ij}, namely,

$$p_{ij} \approx \frac{s_{ij}}{s_{ij} + s_{ji}}, \tag{2.2}$$

where s_{ij} is the number of points scored by T_i against T_j. This is the approach we use in this paper, and accordingly treat Bradley-Terry as a σ-type rating system only.

Building on this model, [3] shows how home-field advantage can also be incorporated in Bradley-Terry, a model which we refer to as *Home-Bias Bradley-Terry*. Assuming we have an away team,

T_a, and a home team, T_h, we define

$$p_{ah} = \frac{r_a}{r_a + \theta r_h} \quad \text{and} \quad p_{ha} = \frac{\theta r_h}{r_a + \theta r_h}, \tag{2.3}$$

where $\theta > 0$ is a parameter corresponding to home-field advantage across the entire NFL, but not necessarily for every single team. When $\theta > 1$ there is a home-field advantage while when $\theta < 1$ corresponds to a general road-field advantage. When using Bradley-Terry methods to find such a θ at the season's end—and finding it as we simultaneously find the ratings of individual teams—we compute that $.9 \leq \theta \leq 1.5$ in every season from 2002–2018. These data are statistically significant in showing that there is a general trend toward home-field advantage in the NFL, and we use this fact in Section 2.5.

2.2 Other rating methods

In the W/L method, each team's rating in round k+1 is its winning percentage over the first k rounds of play, that is, the total number of wins divided by the total number of games played. For bookkeeping purposes, we define a tie to count as one-half of a win. The impetus for using W/L as one of our ranking methods comes from [7] in which he claims that any high-powered, computer-aided technique for predicting games should at least do better than just looking at the team with a better win percentage and picking that team. Note that Easterbrook adopts the convention of choosing the home team to win if the win percentages are identical; accordingly, in the extremely rare instance where our other numerical rating systems give exactly the same win probability to both teams in a contest, we choose that the said method picks the home team. Note that the W/L method is only an ω-type rating.

For our deterministic models we use the methods of [8] and [9].[2] The Massey method estimates that the difference in ratings is the expected point differential between two teams. For example, if after k rounds of play T_i and T_j have Massey ratings of $r_i^{(k)} = -12$ and $r_j^{(k)} = 15$, respectively, then we would expect T_j to beat T_i by $15 - (-12) = 27$ points if they were to play in round k+1. The Colley method finds rating for teams based on their *strength of schedule*: to compute $r_i^{(k)}$, Colley takes into account the number of wins of T_i and also looks at the teams defeated by T_i, taking into account how many wins they have. With this in mind, Massey is a σ-type model by design, although we modify it to also make a corresponding ω-type, whereas Colley is strictly an ω-type model.

For our Markov, or network, methods we use the model in [6], the Biased Voter model developed in [10], the PageRank model in [11], and the Oracle model developed in [12]. All of these methods may be derived from a directed network associated to the game outcomes where

i. each of our N teams is represented by a node,

ii. a directed edge exists from node i to node j if T_i has lost to T_j, and

iii. each of these directed edges has some transitional probability assigned to it.

One fundamental difference between each of our four chosen methods is the way in which the transitional probabilities are assigned, but note that for each of these methods we can make both σ-type and ω-type ratings. Second, the Biased Voter, PageRank, and Oracle methods allow for transitions between nodes besides just those that have played each other: Biased Voter allows for

[2]Before the NCAA switched to a committee to select its four playoff teams, they used the BCS to select teams to certain postseason games, including the national championship game. Both the Massey and Colley methods were included in the BCS's formulation; however, Massey's method is proprietary, while Colley's is publicly available. Accordingly, the Massey method we use here is the one that he uses in his undergraduate thesis.

an edge from a node to itself; PageRank allows transitioning from any node to another node via a *teleportation matrix*; the Oracle model introduces an N+1-st node, called the *Oracle node*, which can be thought of as an N+1-st team that has both lost to and beaten every team.

2.3 Prediction probabilities

We have discussed how to assign ratings to the teams of our league, and in this section, we define how to use those ratings to assign probabilities to individual matches. Given a match between teams T_i and T_j, we let $p_{ij}^{(k+1)}$ denote the *probability that T_i beats T_j in the k+1-st round of the tournament*. Since we would like to include home-field advantage in our predictions, suppose we have an away team, T_a, and a home team, T_h.

2.3.1 Bradley-terry-type probabilities

Using a modified Home-Bias Bradley-Terry-type estimate for the observed probability, we define

$$p_{ah}^{(k+1)} = \frac{r_a^{(k)}}{r_a^{(k)} + \theta_k r_h^{(k)}} \quad \text{and} \quad p_{ha}^{(k+1)} = \frac{\theta_k r_h^{(k)}}{r_a^{(k)} + \theta_k r_h^{(k)}}, \tag{2.4}$$

where $r_a^{(k)}$ and $r_h^{(k)}$ are computed based on a given rating model, and $\theta_k > 0$ is the home-field factor computed after week k. In Section 2.5 we discuss both the computation of this parameter and the method of choosing θ_k, which is crucial to the point of the paper. We note here that defining our probability in this way requires all ratings to be nonnegative (and at least one of the ratings to be positive); however, Massey's ratings do not satisfy this criterion. Accordingly, we use a modified Massey rating, where we normalize the ratings so that $r_i^{(k)} \in [0, 1]$ by considering that each computed rating, $\hat{r}_i^{(k)}$, is an observation based on a normal distribution, X_{Ma}, with mean zero and standard deviation to be the sample standard deviation. Thus, the Massey rating we use in our analysis is defined to be $r_i^{(k)} = P\left(X_{Ma} \leq \hat{r}_i^{(k)}\right)$, and we discuss this choice of normalization in Section 4.

2.3.2 Logistic regression

One of the standard machine learning models is the logistic regression. Here we simply consider the win/loss outcome of each game as a binary variable where 1 is given for a win and 0 for a loss. Next, we estimate $p_{ah}^{(k+1)}$ using the logit function as a linear regression on the ratings of the teams at round k as follows:

$$\text{logit}(p_{ah}^{(k+1)}) = \log\left(\frac{p_{ah}^{(k+1)}}{1 - p_{ah}^{(k+1)}}\right) = \beta_{0,k} + \beta_{1,k} r_a^{(k)} + \beta_{2,k} r_h^{(k)}.$$

Solving for $p_{ah}^{(k+1)}$ gives that

$$p_{ah}^{(k+1)} = \frac{1}{1 + e^{-\left(\beta_{0,k} + \beta_{1,k} r_a^{(k)} + \beta_{2,k} r_h^{(k)}\right)}}. \tag{2.5}$$

Now that we have various methods of assigning probabilities to each of our predictions, we move to a discussion of how we make, and subsequently assess, those predictions.

2.4 Quantity vs. Quality

To begin, using any of the given methods to assign the value $p_{ah}^{(k+1)}$, we make a *prediction* that T_h will beat T_a in week k+1 whenever $p_{ha}^{(k+1)} \geq 0.5$, and this prediction will be correct if T_h actually wins the

game. Thus, when we speak of "quantity," we are simply referring to prediction accuracy. Formally, suppose we are given a collection of predictions for some m events, $\Pi = (\pi_1, \pi_2, \ldots, \pi_m)$. Then the *accuracy* of Π, $Acc(\Pi)$, is the number of correct predictions divided by the total predictions. Now suppose $T = (\tau_1, \tau_2, \ldots, \tau_m)$ is a collection of predictions for the same m events as Π. We say that Π has greater accuracy than T whenever $Acc(\Pi) > Acc(T)$. We want to hold this up against the notion of quality, and while accuracy and quality are certainly not mutually exclusive ideas, there are indeed different.

There are a variety of statistical metrics, called *scoring rules*, that are used to assess the quality of a prediction, and [13] provides an extensive discussion on how to select such metrics. In fact, starting with [14], one can argue that a forecaster can improve a score by modifying the predictions in light of the metric. Therefore, one must find a *proper scoring function* to ensure the highest quality of prediction is attained. More precisely, proper scoring functions are metrics whose expected values are optimized if the forecasted probability is the true event probability. Although there exist many proper scoring functions, only a small number of them are employed in practice, and for a formal discussion about a multitude of scoring functions we refer the reader to the work in [15] and [16]. For our purposes, we only consider the Quadratic Score function (QS). Formally, the QS of a given event is a numerical value given by $2p_{obs} - \|\mathbf{p}\|^2$ where p_{obs} is the probability of the observed outcome and $\mathbf{p}$ is the vector containing the probabilities for all of the possible outcomes. When applied to an event that has only two predicted outcomes, such as an NFL contest, [17] normalize the QS of an individual prediction, π, as $QS(\pi) = 1 - 4(1 - p_{win})^2$, where p_{win} is the probability assigned to the winning team in the prediction π, and we adopt this as our method of assigning a QS value to any individual prediction. Then, given a collection of predictions for some m events, $\Pi = (\pi_1, \pi_2, \ldots, \pi_m)$, the overall QS of Π is

$$QS(\Pi) := \sum_{i=1}^{m} QS(\pi_i).$$

For instance, in our motivating example from Section 1, Player 1 attains a QS of

$$[1 - 4(1 - 0.6)^2] + [1 - 4(1 - 0.55)^2] + [1 - 4(1 - 0.51)^2] = 0.5896,$$

while Player 2 attains a QS of

$$[1 - 4(1 - 0.9)^2] + [1 - 4(1 - 0.8)^2] + [1 - 4(1 - 0.45)^2] = 1.59.$$

To be clear, when Player 2 picks F to beat E with a confidence of 55%, the QS computes this as Player 2 picking the winner, i.e., E, with a confidence of $p_{win} = 0.45$.

Given this normalization of QS, we see that an individual prediction with higher QS is better, and thus a higher total QS score across a set of predictions is desirable. Therefore, given two individual predictions, π and τ, we say that π is a *higher quality* prediction than τ if $QS(\pi) > QS(\tau)$. Suppose further that we are given two collections of predictions for the same m events, $\Pi = (\pi_1, \pi_2, \ldots, \pi_m)$ and $T = (\tau_1, \tau_2, \ldots, \tau_m)$. We say that Π has greater quality than T whenever $QS(\Pi) > QS(T)$. With these definitions in hand, our reason for choosing QS as our one scoring rule is forthright: in hindsight, Home-Bias Bradley-Terry—i.e., the model from which we base our method of determining outcome probabilities—maximizes the QS. It is important to be explicit here in also saying that we are assessing foresight prediction quality, and so there is no *a priori* reason to assume that any version of Bradley-Terry should attain the highest foresight prediction QS score when compared to other models.

2.5 Choosing θ

We introduced a home-field parameter, θ_k, into the mix in Section 2.3.1, and we now discuss two ways in which we compute this value: optimizing for accuracy and optimizing for quality. In both

ways, suppose we have a fixed ratings model M, and after round k we compute the ratings vector for M in the usual way described in Section 2. We then pick a value $0.5 \leq \theta \leq 2.0$ to create a new set of predictions $\Pi^{\theta}_{M,k}$ for *every contest that has already been predicted from rounds 1 though k*, where all of these predictions are of the form

$$\hat{p}^{(k)}_{ah} = \frac{r^{(k)}_a}{r^{(k)}_a + \theta r^{(k)}_h} \quad \text{and} \quad \hat{p}^{(k)}_{ha} = \frac{\theta r^{(k)}_h}{r^{(k)}_a + \theta r^{(k)}_h},$$

for some home team, T_h, and away team, T_a. We define

$${}_{Acc}\theta_k = \min_{\theta \in [0.5,2.0]} \{\theta \text{ maximizes } Acc(\Pi^{\theta}_{M,k})\},$$

and then we set our k+1-st week's prediction probabilities to be

$${}_{Acc}p^{(k+1)}_{ah} = \frac{r^{(k)}_a}{r^{(k)}_a +_{Acc} \theta_k r^{(k)}_h} \quad \text{and} \quad {}_{Acc}p^{(k+1)}_{ha} = \frac{{}_{Acc}\theta_k r^{(k)}_h}{r^{(k)}_a +_{Acc} \theta_k r^{(k)}_h}. \tag{2.6}$$

Similarly, we define

$${}_{QS}\theta_k = \min_{\theta \in [0.5,2.0]} \{\theta \text{ maximizes } QS(\Pi^{\theta}_{M,k})\},$$

and then we set our k+1-st week's prediction probabilities to be

$${}_{QS}p^{(k+1)}_{ah} = \frac{r^{(k)}_a}{r^{(k)}_a +_{QS} \theta_k r^{(k)}_h} \quad \text{and} \quad {}_{QS}p^{(k+1)}_{ha} = \frac{{}_{QS}\theta_k r^{(k)}_h}{r^{(k)}_a +_{QS} \theta_k r^{(k)}_h}. \tag{2.7}$$

3 Results

As mentioned at the end of Section 1, the main goal of this paper is to see how adjusting for accuracy or quality affects overall pick accuracy. To that end, we offer four sets of results, broken up by whether or not we are using a ω- or σ-type model, and whether or not we are making foresight predictions for Weeks 4–17 of the NFL season or Weeks 11–17 of the NFL season. We must wait until Week 4 so that the underlying network is strongly connected and all models are capable of making predictions. We also consider predictions starting at Week 11, as this corresponds to the time when about 70% of the games have been played and all models have significant training data. Again, we reiterate here that we are *less* interested in the actual prediction percentages and *more* interested in whether or not there is a positive change in the prediction percentages.

Figures 3, 3, 3, and 3 each contain four types of prediction percentages: Usual, Logistic, Quantity, Quality. The Usual prediction method uses probabilities defined by Equation (2.4) with $\theta_k = 1$ for all k; the Logistic prediction method uses probabilities defined by Equation (2.5); the Quantity prediction method uses probabilities defined by Equation (2.6); the Quality prediction method uses probabilities defined by Equation (2.7). All of these figures include the W/L model, denoted as WinP, because as referenced earlier, computer models should be able to do better than this method of prediction by [7], and we do not adjust for scores in this model, so the W/L percentages shown in Figures 3 and 3 are indeed the same as in Figures 3 and 3, respectively. Of course, we see from Figure 3 that under the Usual probabilities and not accounting for game scores, the W/L method performs better than our computer-based models.

In addition, when we incorporate home-field advantage via Equation (2.3), we do so for Bradley-Terry by implementing Keener's scoring approximation from Equation (2.2). Accordingly, we evaluate Bradley-Terry as a σ-type model only. Further, implementing Home-Bias Bradley-Terry, described in Section 2.1, at the end of each week produces the value of θ_k which maximizes in *hindsight* the QS, but this method is different from our own for other models. Specifically, when

we compute ${}_{QS}\theta_k$, we do so after first producing a ranking of our teams, but in the Home-Bias model, the team ratings and the θ are found simultaneously using MLE. For this reason, the final bar corresponding to Bradley-Terry in each of Figures 3, 3, 3, and 3 corresponds to this Home-Bias model.

Finally, for each of the overall percentage figures we provide the accompanying Figures 3, 3, 3, and 3, which show the percentage change versus the Usual method when using the Logistic, Quantity, Quality, or Home-Bias method. For example, the ω-type Massey prediction percentage for Weeks 4–17 is 61.6% using the Usual method and 62.8% using the Logistic method (both seen in Figure 3), so the corresponding Logistic value in Figure 3 represents the difference: $62.8 - 61.6 = 1.2\%$.

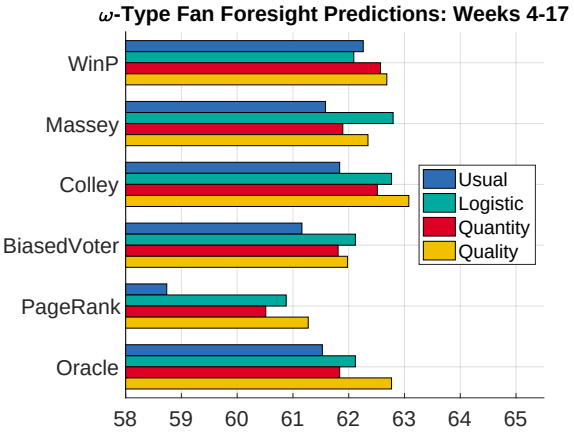

Fig. 1. ω-Type Fan Foresight, Weeks 4–17

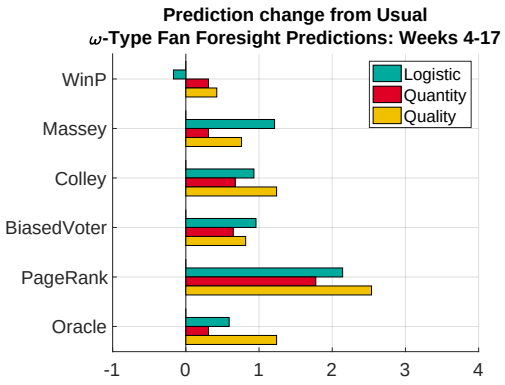

Fig. 2. ω-Type Fan Foresight, Weeks 4–17

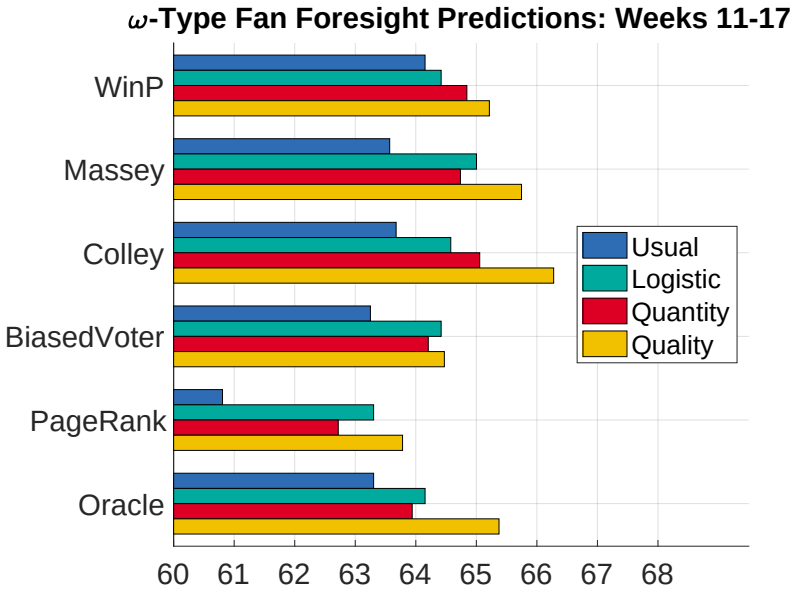

Fig. 3. ω-Type Fan Foresight, Weeks 11–17

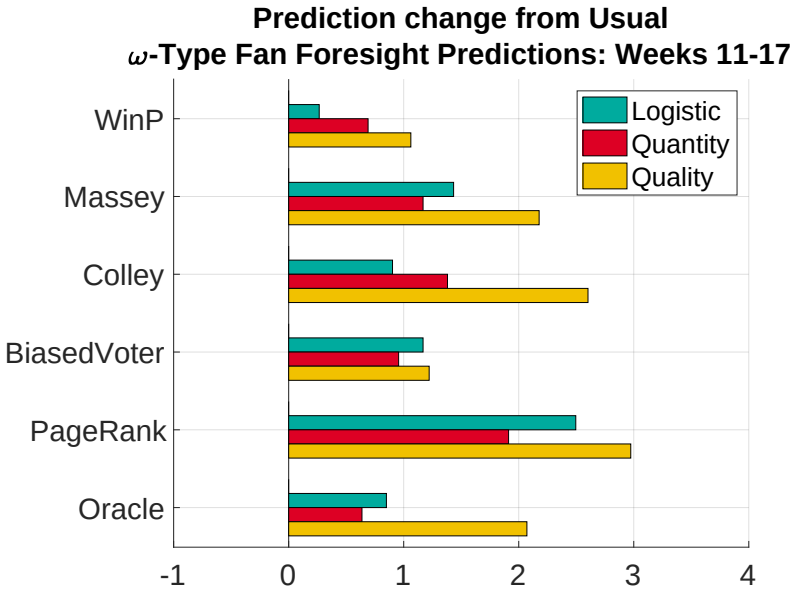

Fig. 4. ω-Type Fan Foresight, Weeks 11–17

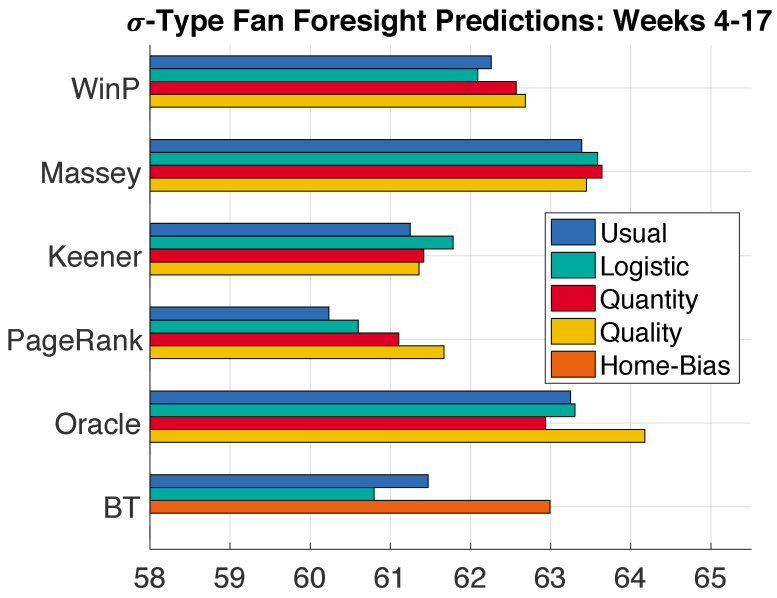

Fig. 5. σ-Type Fan Foresight, Weeks 4–17

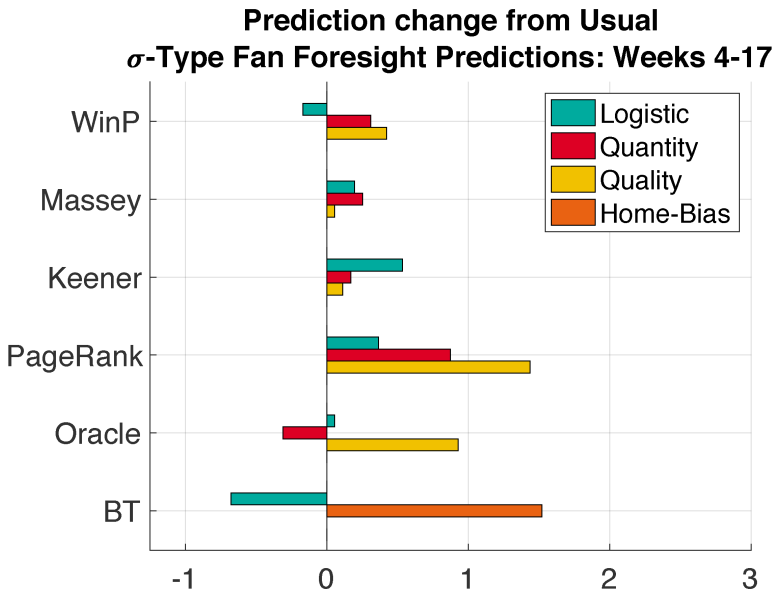

Fig. 6. σ-Type Fan Foresight, Weeks 4–17

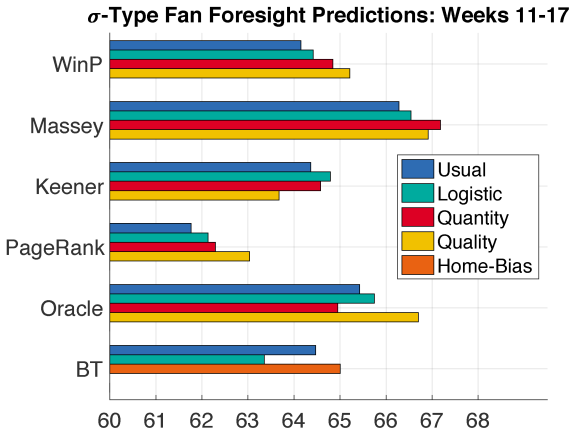

Fig. 7. σ-Type Fan Foresight, Weeks 11–17

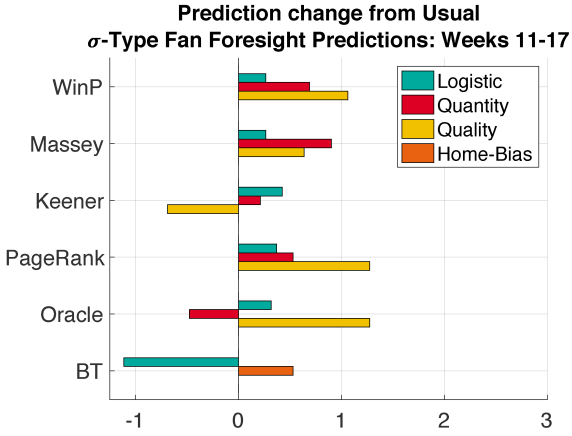

Fig. 8. σ-Type Fan Foresight, Weeks 11–17

4 Discussion

The choice of the home-field factor θ is done in two ways: optimize the quantity of correct games or optimize the quality of the probabilities associated to the winner of the games. In both cases, the overall results show that introducing a home-field factor to make a prediction for games outcomes consistently improves accuracy, and for comparison, we also analyzed how a standard method like the logistic regression would improve accuracy.

The overall results show that introducing a home-field factor to make a prediction for games outcomes consistently improves accuracy. The choice of the home-field factor θ is done to optimize the quantity of correct games or the quality of the probabilities associated to the winner of the games. For comparison, we also analyzed how a standard method like the logistic regression would improve accuracy.

With the exception of the ω-type Keener and Oracle models, all methods improved by using either logistic regression or a θ-optimization. Also, using logistic regression is best only in the Weeks 4–17 ω-type Massey and Biased Voter predictions, and in the former case, this is the type of prediction where we most drastically alter the original Massey method. In particular, we see that all of the Massey percentages are much better in Figures 3, 3, and 3 than they are in Figure 3. Looking to the σ-type models, Massey is improved by all parameterizations, whereas Keener, PageRank (Score), and the Oracle are not. In the case of Keener, optimizing θ for accuracy performs best, whereas optimizing for quality performs best in both the Oracle and PageRank cases.

4.1 Other notes

When thinking about the normalization of the Massey rating vector, we observe that this does not change the relative ranking order, that is, if T_i would be favored against T_j it will continue to be the favorite after the normalization. Of course, this does lose information that the original Massey rating contains. Still, we are only interested in whether or not we can somehow improve our predictions by picking an appropriate θ. This method of normalization is not meant to diminish the pick accuracy of the original Massey method, but any method of making the Massey ratings nonnegative has some aspects which are not desirable with regard to making, and assigning probabilities to, predictions.

Numerically speaking, we also notice variations week-to-week in the values of ${}_{Acc}\theta_k$ and ${}_{QS}\theta_k$, so that knowing a θ value from week k does not help in predicting what the subsequent θ value will be in week k+1, making it necessary to recompute a new θ value in each week of predictions. To better illustrate this, Figures 4.1 and 4.1 show the 2018 weekly computed ${}_{Acc}\theta_k$ and ${}_{QS}\theta_k$ values, respectively, for $4 \leq k \leq 17$ (note that the vertical axes are scaled are differently for easier visualization). We only provide this data for the 2018 NFL season as this is the typical behavior of these parameters across all seasons.

Our analysis can be further improved by considering a θ value for each team. This would be computationally more demanding, but it is certainly reasonable to expect home-field advantage has different effects for different teams.

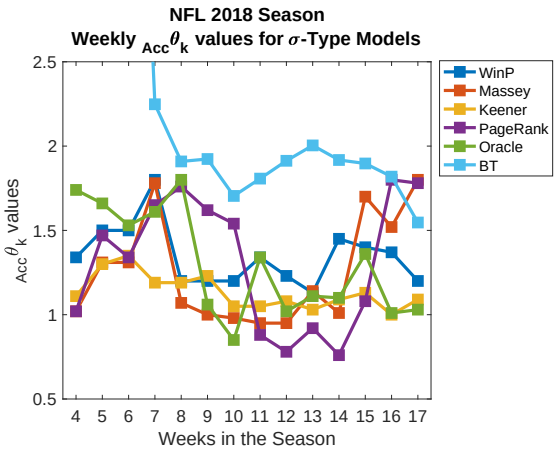

Fig. 9. ${}_{Acc}\theta_k$ for Weeks 4–17 of the 2018 NFL Season

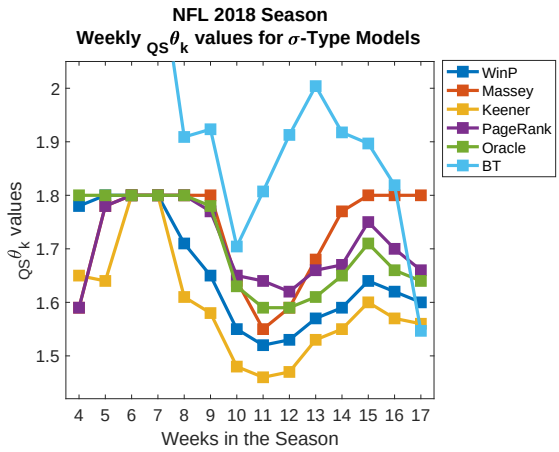

Fig. 10. $_{QS}\theta_k$ **for Weeks 4–17 of the 2018 NFL Season**

5 Conclusion

We showed that choosing a home-field factor to optimize for accuracy or quality affects overall pick accuracy of a ranking method. The increase from a usual prediction method is at times over 2%. As we observe that there were 3548 games predicted between Weeks 4 and Week 17 and 1883 games between Week 11 and 17. The increase in accuracy translates to correctly picking two extra games each season and is done without any added information besides knowing the scores and which team played at home. Moreover, the fact that some methods are significantly more improved by parameterizing with an eye toward optimizing the quality rather than quantity shows that the focus when making a prediction should be on quality alone, and that quantity will naturally follow.

In the future, we expect to analyze other sports leagues to see if the focus on quality over quantity is robust across different leagues. We also hope to find a way to utilize the previous season's θ in order to make predictions for the early games in the next season.

Acknowledgment

The authors thank the reviewers for their helpful comments that have improved the final manuscript.

Competing Interests

Authors have declared that no competing interests exist.

References

[1] Langville AN, Meyer CD. Who's #1?: The science of rating and ranking. Princeton University Press; 2012.

[2] Bradley RA, Terry ME. Rank analysis of incomplete block designs. I. The method of paired comparisons. Biometrika. 1952;39:324-345.

[3] Agresti A. Categorical data analysis: Wiley series in probability and statistics. Wiley-Interscience, 2nd edition; 2002.

[4] Hunter DR. MM algorithms for generalized Bradley-Terry Models. The Annals of Statistics. 2004;32(1):384-406.

[5] Jr. Ford LR. Solution of a ranking problem from binary comparisons. Amer. Math. Monthly. 1957;64(8):28-33.

[6] Keener JP. The Perron-Frobenius theorem and the ranking of football teams. SIAM Review. 1993;35(1):80-93.

[7] Easterbrook G. Time to look back on some horrible predictions; 2008.
Available: http://www.espn.com/espn/page2/story?page=easterbrook/080212
(Accessed on 1 July 2019)

[8] Massey K. Statistical models applied to the rating of sports teams. Bluefield College Bachelor's Thesis; 1997.

[9] Colley W. Colley's bias free college football ranking method: The colley matrix explained; 2002.
Available: http://www.colleyrankings.com/matrate.pdf.
(Accessed on 1 July 2019)

[10] Callaghan T, Mucha PJ, Porter MA. Random walker ranking for NCAA division I-A football. Amer. Math. Monthly. 2007;114(9):761-777.

[11] Brin S, Page L. The anatomy of a large-scale hypertextual web search engine. Comput. Netw. ISDN Syst. 1998;30:107-117.

[12] Balreira EC, Miceli BK, Tegtmeyer T. An oracle method to predict NFL games. Journal of Quantitative Analysis in Sports. 2014;10:1-14.

[13] Merkle EC, Steyvers M. Choosing a strictly proper scoring rule. Decision Analysis. 2013;10(4):292-304.

[14] Brier GW. Verification of forecasts expressed in terms of probability. Weather Rev. 1950;78:1-3.

[15] Gneiting T, Raftery AE. Strictly proper scoring rules, prediction and estimation. Journal of the American Statistical Association. 2007;102(477):359-378.

[16] Servan-Schreiber E, Wolfers J, Pennock DM, Galebach B. Prediction markets: Does money matter?. Electronic Markets. 2004;14:243-251.

[17] Chen Y, Chu CH, Mullen T, Pennock DM. Information markets vs. opinion pools: An empirical comparison. In: Proceedings of the 6th ACM conference on Electronic commerce. 2005;58-67.

Peer-review history:
The peer review history for this paper can be accessed here (Please copy paste the total link in your browser address bar)
http://www.sdiarticle4.com/review-history/52105