PREDICTIVE MODELS FOR GAME OUTCOMES IN WOMEN'S LACROSSE

Michael Scott Brown

The Graduate School, University of Maryland University College, Adelphi, Maryland, United States

ABSTRACT

This research presents a predictive model for determining the game outcome of a Women's (Female) Lacrosse game. This is important to coaches regardless of if their team appears to be winning or losing the game. Coaches make decisions throughout the game based upon the belief that they are winning or losing. The model is a Logistic Regression model and can be used with very little data from a game: time remaining and difference between the scores. This could be a valuable tool to coaches that can be used during the game. It is more than 89% accurate. Data used in this research comes from direct matchup games between BigTen Women's Lacrosse teams. The win probability equations, including coefficients, are presented.

KEYWORDS

End of Game Prediction, Win Probability, Lacrosse, Women's Lacrosse, Female Lacrosse, Logistic Regression.

1. INTRODUCTION

Lacrosse is a game invented by the Native Americans in North America. It is played between two teams on a large field. Each team normally has between 10 and 12 players on the field and scores points by getting a small ball into the opposing team's goal which are located on opposite ends of the field. Players cannot touch the ball, rather they use a special stick with a net on the end of it. Using the stick, they can scoop the ball into the net allowing them to pass, catch and shoot the ball at the goal. The team with the most goals at the end of the game is the winner. As with most sports there are many other rules, but this is the basic concept of the game [1].

The first lacrosse games were witnessed and recorded by the Europeans in the 1600's. This allowed it to spread to Europe. In 1890 a women's version of it was invented in Scotland by adapting the basic rules to reduce physical contact between players [2]. Today there are a few variations of the rules for Women's' Lacrosse [1, 3].

Lacrosse is growing in popularity. According to USLacrosse participation in the game has increased 225% over a 15-year period [4] ending in 2016. Women's Lacrosse is played by over 300,000 women and girls across the United States [4]and this number is increasing yearly. Additionally, it is played in some other countries in Europe, other English speaking countries and very large countries like China.

DOI : 10.5121/mathsj.2019.6101

1.1. Important to Know Game Outcome

It would be beneficial to know the probability of the outcome of the game. There are a number of decisions that coaches make throughout the game that depend upon knowing the probability of the outcome of the game. As in many other sports game outcome predictions is a problem that research needs to address.

If a coach believes the probability that their team will win the game, there are a number of decisions that can be made. A coach may slow down the tempo of the game by having the players perform safe activities that decrease the chance that the other team may get the ball. One example of this is to have players pass the ball back and forth away from the goal. Opposing players often stay close to their goal to prevent the other team from scoring. This consumes time and gives the other team less chances to change the outcome of the game. The team's goal is to consume time and not necessarily to score more points. Coaches can change the tempo of the game by signaling from the side lines.

Coaches may believe that their probability of winning is so high that the coach takes out the better players from the game and substitutes younger, lesser skilled players to give them more experience. Someday these substitutes might be needed in a game and this gives them valuable experience in a low pressure environment. This is often referred to as putting in your second stringers, which is a term used for lesser skilled players.

If a coach believes that the probability of losing the game is high, there are decisions that can be made to try and change the outcome of the game. A coach might call a time-out, which stops the game for a short period of time. This gives their players a chance to rest and could change the momentum of the game. The time out also gives the coach an opportunity to give players detailed instructions.

A coach believing that a loss is likely could signal to have their players play more aggressively. This could increase the number of favorable turn-overs of the ball, which could reverse the outcome of the game. Typically, more aggressive play gives the other team greater chances to score. But in a game that the team is likely to lose, this might be worth the risk.

In some rare cases when a coach believes that there is no chance of winning a game, they might put in their second string players. This increases the chance of losing, but gives second string players experience which is important for future growth of the team. So, they sacrifice the game to gain experience for their second string players.

A game outcome predictive model would benefit coaches during the game. It would aide in the decision making processes that take place throughout the game. This justifies research in this area.

2. LITERATURE REVIEW

2.1. Game Outcome Research

There has been little research in the field of predictive sports game outcome. One notable exception is McFarlane's [22] paper on predicting the outcome of Basketball games. But it is limited to only the last 3 minutes of the game. This logistic regression [5] model is beneficial for coaches to determine if they should intentionally foul the other team in hopes of getting the ball back. The model is also useful in determining if your players should make harder 3-point shots over easier 2-point shots.

One area of game outcome prediction centers around sports betting. While many papers attempt to address this problem, all of them only look at pre-game data [6-8]. This is because you can only bet on a sport before it begins. While this research addresses problems arounds sports betting it is not useful in real-time decision-making throughout the game.

2.2. Player Decision Making

Another area of research attempts to address problems with player decision making. Players are faced with numerous decisions throughout a game. In fact, many strategies in sports try to set up paradoxical decision situations for the opposing players. So, no matter which decision is made it ultimately is wrong. In other strategies players are faced with situations in which the obvious decision is wrong. Much research has been done in this area ranging from speed and accuracy of decision making [9]to operational blindness [10]and many others [11-12].

2.3. Women's Lacrosse

There is very little research published on Women's Lacrosse or even Lacrosse in general. One area of research focuses on player practice and game injuries [13-14]. While this is important research it is not applicable to coaches' decision making during the game. Another area of research addresses the problem of player selection [15-16]. Player selection is typically done before the seasons starts and does not address problems with decision making during a game. Women's Lacrosse is a complex game but no research exists to assist coaches in decision making during a game.

3. METHODOLOGY

3.1. Data

The data for this research comes from the NCAA [17] Play-by-play data. The data comes in the form of an HTML table for Period 1, Period 2 and in some cases Overtime. The table has the time, in minutes and seconds, until the period ends. There is a home team column with free form text about an event in the game, followed by a score column and finally a free form text column for the visiting team. Scores are shown as home team score, hyphen, visiting team score.

Because some games are mismatches, this research only used games between two BigTen teams. The BigTen is a collegiate sports conference located in the Northeast and Great Lakes area of the United States made up of very large colleges. There are seven BigTen Women's Lacrosse programs: Maryland, Johns Hopkins, Northwestern, Rutgers, Penn State, Michigan and Ohio State. Data comes from the 2016-2017 season that consists of games played in the Spring 2017. This was the most recent full year available at the time of this research and includes 21 games. A new data point is created each time that the score changes. In the 21 games there were 517 score changes.

Data from the NCAA site was preprocessed into an easy to work with format. Time was converted into the number of seconds before regulation time expires, Overtime is considered negative time. There are two 30 minute periods in Lacrosse, so the beginning time was 3,600 seconds until the game ends. This data was normalized making time a value of 1.0 when the game begins and 0.0 when the game ends. The Momentum for each team is also computed. The Momentum of a team is defined by the slope of the current score compared to the two prior scores.

Here is an example of computing Momentum. On March 18, 2017 Michigan played Ohio State. At 25:05 into the first period Ohio State Scored. At 19:43 Michigan scored. These scores are shown in Table 1. To compute Momentum at 19:43, record 3, to the time and scores in record 1. The Momentum for Michigan at 19:43 is (1 – 0) * 100/ (3600 – 3013) = 0.17. The Momentum for Ohio State is also 0.17.

Table 1. Example Data for Computing Momentum

Record #	Time	Time in Second to end of game	Michigan	Ohio State
1	Period 1, 30:00	3,600	0	0
2	Period 1, 25:05	3,305	0	1
3	Period 1, 19:43	3,013	1	1

The dependent variable is an additional column, which is the outcome of the game from the first team's perspective. Since all of the NCAA data the first team is the home team and that might be a factor in the outcomes, the pre-processing program for the data randomly selects a team to the be the first team. The values for game outcome are "won" and "lost". The dataset consists of 516 records.

3.2. Algorithm

Regression Analysis is a technique of determining coefficient values for an equation that produces the best fit to a set of data. Common methods are linear regression and polynomial regression[18]. This research used logistic regression [19-20]which is ideal for producing probabilities of an event. All values of logistic regression range between 0% and 100%. This method is classified as Supervised Learning algorithms [21].

Logistic Regression is a statistical technique that generates an equation that maps the independent variables to the probability of each of the discrete values for the dependent variable [19]. In the case of this research the independent variable is defined as the ratio of difference between scores, *S*, from the first team's perspective to normalized time remaining, *NT*.

The dependent variable is the game outcome which is Won or Lost. The algorithm determine the coefficient and intercept to minimize the error between the model and the training data. It will produce equation 1.

$$P(Win) = \frac{1}{1+\exp\left(-(B_0+B_1(S/NT))\right)} \quad (1)$$

Logistic Regression will find the values for $B_0$and B_1 that produce the least squared error. The value of a Win was selected arbitrarily for the algorithm. Computing the probability of losing is shown in equation 2.

$$P(Loss) = 1 - P(Win) \quad (2)$$

A second Logistic Regression equation was used that include each team's momentum. In this equation the momentum of Team 1 is M1 and the momentum of Team 2 is M2. Equation 3 shows the logistic regression equation that includes momentum.

$$P(Win) = \frac{1}{1+\exp\left(-(B_0+B_1\left(\frac{S}{NT}\right)+B_2\left(\frac{M_1}{NT}\right)+B_3\left(\frac{M_2}{Nt}\right))\right)} \quad (3)$$

3.3. Format of Results

This research used 75% of the 516 records as training data and the remaining 25% as test data. These records were selected randomly across the data file. The results will include the percent of test data records correctly classified, average precision, average recall and FScore. The percent of test data records correctly classified is self-explanatory.

Precision and recall are often used as measurements of predictive models. The equation for precision is shown in equation 4; recall is shown in equation 5.

$$precision = \frac{true\ positive}{true\ positive + false\ positive} \tag{4}$$

$$recall = \frac{true\ positive}{true\ positive + false\ negaitve} \tag{5}$$

When attempting to predict an outcome, without loss of generality a Win, four numbers can be computed. True positive means the prediction is Positive, a Win, and the outcome was positive, a Win. False positive means the prediction is Positive, a Win, and the outcome was negative, a Loss. True negative means the prediction is Negative, a Loss, and the outcome was negative, a Loss. False negative means the prediction is Negative, a Loss, and the outcome was negative, a Loss. With these four number precision and recall can be computed. In this research you can try to predict a Win or a Loss, the average of the two is reported.

The F Score takes into account the precision and recall. It is the harmonic average of the precision and recall and is shown in equation 6.

$$FScore = \frac{2}{\frac{1}{precision} + \frac{1}{recall}} \tag{6}$$

In cases of precision, recall and F Score all scores range between 0 and 1 and higher numbers indicate a better predictive model.

Figure 1 shows a line chart of the probability of winning based upon the amount of time left in the game. The chart shows the probability lines for all possible point differences ranging from being ahead by 4 points to being behind by 4 points. The values for time decrease along the X-axis, starting at 3,600 and going to 0. This allows the reader to look at the chart from left to right as the game progresses.

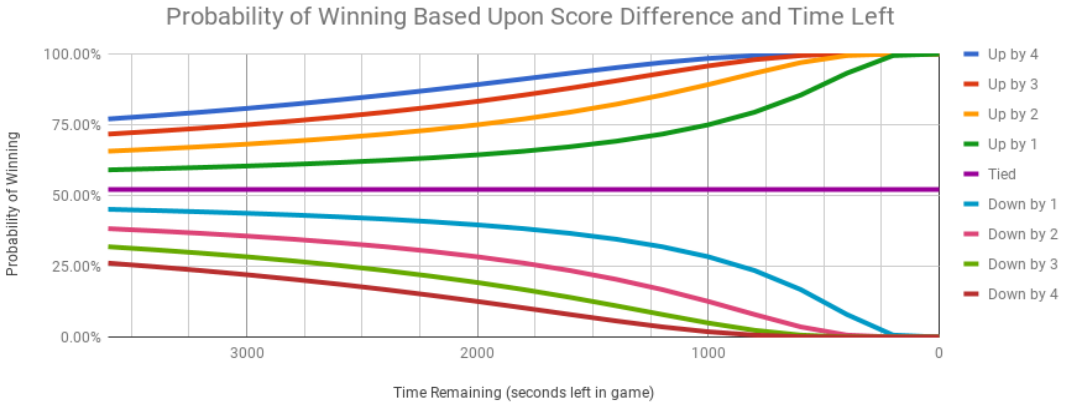

Figure 1. Probability of Winning Based Upon Score Difference and Time Left

4. RESULTS

The Logistic Regression algorithm produced B_0 and B_1 to be 0.085 and 0.2812 respectively, giving the final equation to be shows in equation 7.

$$P(Win) = \frac{1}{1+\exp\left(-(.85+0.2812(S/NT))\right)} \tag{7}$$

Of the 25% of data records used for testing, the algorithm correctly identified the correct outcome of the game 89.15% of the time. The Precision, Recall and F Score can be seen in Table 2.

For the equation that included momentum the following coefficient values were found. B_0 is 0.1149; B_1 is 0.2821; B_2 is -0.025 and B_3 is -0.0111. Equation 8 shows the final equation.

$$P(Win) = \frac{1}{1+\exp\left(-(0.1148+0.2821\left(\frac{S}{NT}\right)+(-0.025)\left(\frac{M_1}{NT}\right)+(-.0111)\left(\frac{M_2}{Nt}\right))\right)} \tag{8}$$

This equation correctly predicted the outcome of the game 89.92% of the time. The Precision, Recall and F Score can also be seen in Table 2.

Table 2. Results for Two Equations

Algorithm	Percent Correctly Classified	Weighted Average Precision	Weighted Average Recall	F Score
S, NT	89.15%	0.893	0.891	0.891
S, NT, M	89.92%	0.899	0.899	0.899

The equation that includes momentum produces a slightly higher accuracy of 89.92% compared to 89.15% without momentum. With sample sizes of N = 129, this produces a p value of 0.8408 using the Chi-Square test. This is not a statistically significant difference. If there is not a statistically significant difference between the two equations the simpler one is preferable, which is equation 7. When programmed equation 7 executes faster because it does not need to calculate momentum.

5. CONCLUSIONS

This research shows that a simple logistic regression model can be very effective at predicting the outcome of a Women's Lacrosse game. This can be a valuable coaching aide during a game. Information from the model can be used by coaches to help secure wins when they are ahead and change the outcome of games when they are behind.

More research needs to be done in this area. Other predictive methods may address this problem better. There are other problems that need to be addressed by future research. Future research questions that should be addressed are:

- How effective are different types of coaching decisions on the outcome of a game?
- When is the ideal time for coaches to make changes in a game?
- How effectively can Lacrosse games be simulated?
- This model was developed using data from College Division 1 games. How does the model change when using data from other lacrosse leagues?

Lacrosse is a game growing in popularity around the world. Even though the rules are simple, the game is very complex. With Recreational, Club, High School, College and National teams competing for wins it is only natural that analytics will eventually become part of the game. Artificial Intelligence and predictive analytics can be valuable tools for coaches and teams.

REFERENCES

[1] NCAA. 2018. 2018 and 2019 NCAA Women's Lacrosse Rules.

[2] Tucker, J. and Yakutchik, M., 2014. Women's lacrosse: a guide for advanced players and coaches. JHU Press.

[3] 2018 US Lacrosse. 2018. 2018 Youth Girls' Rulebook: Official Rules for Girls' Lacrosse.

[4] USLacrosse. 2016. 2016 Participation Survey. USLacrosse.org.

[5] Diez, D. M., Barr, C. D. and Cetinkaya-Rundel, M. 2015. OpenIntro Statistics, Third Edition.

[6] Feddersen, A., Humphreys, B.R. and Soebbing, B.P., 2017. Sentiment bias and asset prices: Evidence from sports betting markets and social media. Economic Inquiry, 55(2), pp.1119-1129.

[7] Kain, K.J. and Logan, T.D., 2014. Are sports betting markets prediction markets? Evidence from a new test. Journal of Sports Economics, 15(1), pp.45-63.

[8] Moskowitz, T.J., 2015. Asset pricing and sports betting.

[9] Frehaigne, J.F., Godbout, P. and Bouthier, D. 2001. The Teaching and Learning of Decision Making in Team Sports. Quest, 52, pp.59-76.

[10] Memmert, D. and Furley, P. 2007. "I Spy With My Little Eye!": Breadth of Attention, Inattentional Blindness, and Tactical Decision Making in Team Sports. Sport Psychology 29, pp. 365-381.

[11] Raab, M. 2003. Decision Making in Sports: Influence of Complexity on Implicit and Explicit Learning. International Journal of Sports and Exercise Psychology, 1, pp. 301-337.

[12] Baker, J. and Cote, J. 2003. Sport-Specific Practice and the Development of Expert Decision-Making in Team Ball Sports. Journal of Applied Sports Psychology, 15, pp. 12-25.

[13] Clark, J.M., Hoshizaki, T.B. and Gilchrist, M.D., 2018. Assessing women's lacrosse head impacts using finite element modelling. Journal of the mechanical behavior of biomedical materials.

[14] Dick, R., Lincoln, A.E., Agel, J., Carter, E.A., Marshall, S.W. and Hinton, R.Y. 2007. Descriptive Epidemiology of Collegiate Women's Lacrosse Injuries: National Collegiate Athletic Association Injury Surveillance System, 1988-1989 Through 2003-2004. Journal of Athletic Training, 42(2), pp. 262-269.

[15] Hoffman, J.R., Ratamess, N.A., Neese, K.L., Ross, R.E., Kang, J., Margrelli, J.F., and Faigenbaum, A.D. 2009. The Journal of Strength and Conditioning Association, 23(5), pp. 1524-1529.

[16] Vescovi, J.D., Brown, T.D. and Murray, T.M. 2007. Descriptive characteristics of NCAA Division I women lacrosse players. Journal of Science and Medicine in Sport, 10, pp 334-340.

[17] NCAA.org.

[18] Jiawei, H., Micheline, K. 2007. Data Mining: Concepts and Techniques. Second Edition.

[19] Cox, D.R., 1958. The regression analysis of binary sequences. Journal of the Royal Statistical Society. Series B (Methodological), pp.215-242.

[20] Le Cessie, S. and Van Houwelingen, J.C., 1992. Ridge estimators in logistic regression. Applied statistics, pp.191-201.

[21] Russell, S.J., Norvig, P., Canny, J.F., Malik, J.M. and Edwards, D.D., 2003. Artificial intelligence: a modern approach (Vol. 2, No. 9). Upper Saddle River: Prentice hall.

[22] McFarlane, P., 2018. Evaluating NBA end-of-game decision-making. Journal of Sports Analytics, (Preprint), pp.1-6.

AUTHORS

Michael Scott Brown is the Program Chair for the Software Engineering Master's at the University of Maryland University College. He holds multiple degrees is Mathematics and Computer Science including a PhD in Computer Science.