COMPARING PREDICTIVE MODELS FOR ENGLISH PREMIER LEAGUE GAMES

A Thesis
by
ZACHARY ANDREWS

Submitted to the Graduate School
Appalachian State University
in partial fulfillment of the requirements for the degree of
MASTER OF SCIENCE

May 2019
Department of Computer Science

COMPARING PREDICTIVE MODELS FOR ENGLISH PREMIER LEAGUE GAMES

A Thesis
by
ZACHARY ANDREWS
May 2019

APPROVED BY:

Rahman Tashakkori, Ph.D. Chariperson, Thesis Committee

James B. Fenwick Jr., Ph.D. Member, Thesis Committee

R . Mitchell Parry, Ph.D. Member, Thesis Committe

Rahman Tashakkori, Ph.D. Chairperson, Department of Computer Science

---

Michael McKenzie, Ph.D. Dean, Cratis D. Williams School of Graduate Studies

Abstract

COMPARING PREDICTIVE MODELS FOR ENGLISH PREMIER LEAGUE GAMES

Zachary Andrews

B.S., Appalachian State University

M.S., Appalachian State University

Chairperson: Rahman Tashakkori, Ph.D.

Data science has become an important aspect of modern day society. While the term was first coined in 1960 by Peter Naur, over the past decade, it has been applied to many different fields, one of which is sports. Over the past years, many ranking methods and rating systems have been developed for different sports; the Massey Ranking method, the Elo-rating system, and the Pomeroy ranking method are just a few examples of such models. However, there has been a lack of research in the area of accurate predictive modeling in soccer. The goal of this thesis is to compare and contrast a set of predictive models for determining the outcome of English Premier League (EPL) games.

Acknowledgements

I would like to thank my thesis advisor, Dr. Tashakkori for his patience, wisdom, and support during my time at Appalachian State University, both as an undergraduate student and graduate student. His willingness to help had a great impact on my success here at Appalachian State University. I would also like to thank my committee members, Dr. Fenwick and Dr. Parry, both of who dedicated much time and effort into helping me succeed in the completion of my thesis and degree. Thanks also to the Department of Computer Science and all of its professors for providing me with the knowledge required to be successful here at Appalachian State University. Finally, I would like to thank my family and friends who have supported me throughout this entire journey.

Contents

Abstract **iv**

Acknowledgements **v**

1 Introduction **1**

1.1 Statement of the Problem 1

1.2 Goals 1

1.3 Overview of the Thesis 2

2 Background **3**

2.1 English Premier League History and Structure 3

2.2 Elo Ratings 4

2.3 Poisson Distribution 5

3 Previous Work **7**

3.1 Applications to Sports 7

3.2 Application of Predictive Models in Soccer 10

3.2.1 Mixed Methods 10

3.2.2 Elo Rating 11

3.2.3 Poisson Models 12

3.2.4 Probit Models 12

3.2.5 Factors in Game Prediction 13

4 Methodology **15**

4.1 Data 15

4.2 Elo and Logistic Regression Model 17

4.2.1 Data 17

4.2.2 The ELO Formula 18

4.2.3 ELO Implementation 19

4.3 Dixon & Coles Poisson Model (DC) 21

4.3.1 The Data for the DC Model 21

4.3.2 DC Model Implementation 22

4.3.3 DC Adjustment 22

4.3.4 The Model 23

5 Results **27**

5.1 Elo & Logistic Regression Model 27

5.2 Dixon & Coles Poisson Model 31

6 Conclusions **35**

6.1 Future Work 36

Bibliography **38**

Vita **42**

Chapter 1

Introduction

1.1 Statement of the Problem

Over the past several years, a variety of models in relation to predictive analysis have become increasingly popular. Some of the applications of these models can be traced back to the early 1900s [1]. Predicting the outcome of sports matches has long been of interest to different groups of people. In recent years, a plethora of different models have been produced in order to aid in the prediction of match outcomes [2]. Over the past decade, these techniques have been applied to many different sports, one of which is soccer. Many people have implemented these different models to determine the outcome of soccer matches, but there is little to no comparison between these models. This thesis will compare and contrast two models for determining the outcomes of English Premier League (EPL) matches.

1.2 Goals

This thesis investigates whether certain models perform better than others and attempts to discover what constitutes a successful versus an unsuccessful model when predicting the outcomes of EPL matches. The models being analyzed in this the-

sis include: an adaptation of ELO ratings with multinomial logistic regression, and an implementation of a Poisson model. These models take into consideration input variations such as recent form of a team, home advantage, and strengths and weaknesses of a team's attack and defense. By comparing and contrasting these models with their varying inputs, this thesis determines which models produce better results when predicting match outcomes.

1.3 Overview of the Thesis

Chapter 2 provides a background of key concepts needed in order to understand the work completed in this thesis. Previous work in the areas of predictive analysis, ranking methods, and rating systems is discussed in Chapter 3. Chapter 4 describes the methods used to obtain the relevant data, as well as the process used to compare the models being analyzed in this thesis. The results from the comparisons with be discussed in Chapter 5. Chapter 6 will discuss the conclusions gathered from this research, as well as future work.

Chapter 2

Background

2.1 English Premier League History and Structure

The English Premier League (EPL), based in England, was established in 1992 and is the top level of a larger English League Soccer System. The levels of the English League System from the weakest to the strongest are as follows: League Two, League One, the Championship, and the EPL. Originally, there were 22 teams in the EPL, however, in 1995 the number of teams dropped to 20. All teams in the EPL were from England until the 2011 season, when two Welsh teams joined the league. EPL membership changes structurally each year because of relegation and promotion. The bottom 3 teams are "relegated" at the end of each season to the lower Championship level, and the top two teams from the Championship earn automatic promotion to the EPL. The third team to be promoted is determined by a four-team playoff between the third through sixth place Championship teams. There are a total of 380 games played in an EPL season, with each of the 20 teams playing one another twice. Teams' positions in the league table, or standings, are determined by the number of points they earn. A win earns a team three points, a draw earns a team one point, and a loss earns a team zero points. Goal difference is used as a tiebreaker if teams finish the

season with the same number of points. There is value to earning a high place in the league table, with the top four teams earning automatic qualification to a much larger league tournament known as the Union of European Football Associations (UEFA) Champions League, which is a tournament between the top teams from all leagues in Europe.

2.2 Elo Ratings

The Elo rating system was first introduced to calculate the various skill levels of chess players. Originally these ratings were only used in chess; however, Elo ratings now have many other real-world applications. Most other real-world applications of the Elo rating system involve the rating of sports teams and players in gaming tournaments, with a few other applications. This rating system has been used in such things as the Bowl Championship Series system in college football, the League of Legends videogame tournament, and soft biometrics. In this thesis, Elo ratings will be applied to the prediction of soccer match outcomes.

Elo ratings are calculated based on the result of a match between two teams. After the initial Elo ratings are calculated, those ratings fluctuate as more games are played. In a basic Elo rating system for soccer, the only factors taken into consideration are who the home and away teams are, and the result of the match. For example, consider a match in which Team A is the home team and has an Elo rating of 1600, and team B is the away team and has an Elo rating of 1400. Both teams risk a certain number of points in the match, say 5% of their total points. This means that Team A is risking 80 points while team B is risking 70 points. If Team A was to win, their updated Elo rating would be 1670 while Team B's updated Elo rating would be 1330. The team with the higher Elo rating is considered the stronger of the two teams, so that team will always risk more points.

In a more sophisticated version of the rating system, one may consider factors such as margin of victory and home-field advantage. One major flaw with Elo ratings attempting to determine the outcome of a soccer match is that the resulting calculations can only provide the percentage for a win by the home team or the away team. There is no direct way to determine the likelihood of a draw with both the basic Elo and goal-based Elo ratings. The solution used for that issue is multinomial logistic regression. In regular logistic regression, the dependent variable is categorical and only allows for two choices for the dependent variable. Multinomial logistic regression is a form of logistic regression in which there are more than two possible choices for the dependent variable, a home win, an away win, and a draw in the case of soccer. Given an initial set of Elo ratings, those ratings can be used to estimate the parameters of a multinomial logistic regression model.

2.3 Poisson Distribution

The Poisson distribution can be used as another model to determine the outcomes of EPL games. The Poisson distribution is a probability distribution in which a number of events occur at fixed intervals of time with a known average rate. Some examples of this distribution are the number of network failures per day, the number of patients arriving at an urgent care center between the hours of 8:00am - 9:00am, and the number of defective products a factory makes in a given day [3]. In the case of soccer, this distribution represents the number of goals scored over the full 90 minutes of a soccer match, assuming goals scored throughout a match are independent of one another. Given the mean number of occurrences in a time period, one can obtain the probability of a certain number of those occurrences taking place. The Poisson distribution formula is defined in (2.1).

$$P(x) = \lambda^x e^{-\lambda}/x! \tag{2.1}$$

λ is the mean number of occurrences, x is the number of occurrences, and e is Eulers number.

Chapter 3

Previous Work

Ranking methods and their applications to various sports have been well documented over the past several years. Predicting the outcomes of matches in sports has long been an interest to the general public, as well as bookmakers. Various ranking methods and rating systems can be used to determine the outcome of a particular game. While the methods described in this chapter are being applied to predicting the outcomes of matches in sports, the same methods can be applied to various other fields, including politics, economics, and medicine. This chapter describes previous work in the area of predictive analysis and its application to sports as well as soccer specifically. In regards to soccer, previous work relating to Elo ratings, Poisson models, probit models, game factors, and various mixed methods are described in the second part of this chapter.

3.1 Applications to Sports

A number of researchers have investigated the application of ranking methods and rating systems to sports. Barrow et al. focused on eight different ranking methods and their application to professional basketball, baseball, college basketball, and college football [4]. With each of the eight methods, two different versions of that

particular method were implemented. The first version took into account win-loss data, while the second version took into account score difference data only. The eight methods compared include: winning percentage, Rating Percentage Index (RPI), least squares pairwise comparison, maximum posterior, Keener's method, PageRank, random walker, and Elo ratings. In order to evaluate the accuracy of the predictions for each of these methods, they used 20-fold cross validation. Overall, they were able to conclude that score difference data provided better predictions than win-loss data.

Rating systems and ranking systems have been around for many years and have evolved over time. Pollard and Stefani compared four such systems for soccer, rugby, and college football [5]. The systems used include FIFA and Elo for soccer, IRB for rugby, and BCS for college football. While there are many problems associated with both Elo ratings and FIFA rankings, adaptations have been made to each to improve their accuracy.

In recent years, monitoring a team's chance of winning throughout a particular game has become a point of interest amongst many sports fans and statisticians. Gill focuses on late-game reversals, a losing team coming from behind to ultimately win a game, in professional basketball, football, and hockey [6]. He develops models for basketball and football with the assumption that the scores of these sports are normally distributed. Gill develops a model for hockey with the assumption that the scores of this sport follow a Poisson distribution. In order to compare the expected versus the observed outcomes, a goodness-of-fit test is used for each of the three models. Each of the three models were accurate when applied to their respective sport.

A Brownian motion model can be used to analyze the changes in a team's chance of winning as a match progresses. This model takes into account the margin in which a home team leads or trails as well as the amount of time left to play in a match. This model was applied to 493 professional basketball matches, and by taking into

account the scores at the end of each quarter, it was determined that the Brownian motion model provided a good fit to the results.

Logistic regression models are common when analyzing the rankings of a certain sport. Lebovic and Sigelman applied the model to college football teams in order to determine the number of positions that a team moves up or down in the rankings from week to week [7]. The results of this model showed that a team is more likely to move up if their win is over a higher ranked opponent, and a team will drop in the rankings at a much greater pace if they lose to a lower ranked team.

There are many cases in which statistical models fail to provide better than reasonable predictions than experts of a certain sport. Boulier et al. compared the game predictions of 496 NFL matches [8]. These predictions were made by both statistical models, and experts of professional football. These predictions were compared to each other and also were compared to the predictions made by the betting line. Some of the variables that were used as input to the statistical models included the records of teams, points scored, yards gained, home field advantage, etc. While both failed to beat the predictions of the betting line, the predictions from the experts proved to be superior to those made by the statistical models.

Another outcome in which the betting market proved to be the best predictor of the results of matches came from a study conducted by Boulier and Stekler in relation to predicting the outcomes of National Football League matches [9]. They used the power scores generated by the New York Times to generate probit regression models. The predictions generated from these models were compared to those generated by models based on the betting market as well as the opinions of sports editors. While the probit regression models performed slightly better than the predictions from sports editors, the models based on the betting market were found to be the best at predicting the outcomes of National Football League matches.

Boulier et al. evaluate the predicting power of National Football League matches

using Cohen's kappa coefficient that results in the level of agreement between two variables, in this case, football experts and statistical systems [10]. By using Cohen's kappa coefficient, it was concluded that there is a higher level of agreement amongst statistical systems as opposed to football experts.

Many important concepts and methods can be taken from the literature. Elo ratings can be used as input to both logistic regression models and ordered probit models. Each of these models can be applied to a plethora of sports, including soccer. It is also important to look closely at the idea of analyzing how a team's chance of winning fluctuates as a match progresses. Various game statistics can be used to analyze this concept, and this idea will be analyzed and applied in this thesis.

3.2 Application of Predictive Models in Soccer

This section discusses a variety of prediction methods and their relation to soccer. A few of the methods and techniques described in this section will be used in this thesis.

3.2.1 Mixed Methods

Two distinct types of prediction methods have been applied to soccer, the result of the match (win, lose, or draw) and the number of goals scored and conceded by both teams during a match. Goddard constructed and analyzed two models for each of these prediction methods [11]. Two bivariate Poisson regression models were used to predict the number of goals scored and conceded by teams during a match, and two ordered probit regression models were used to predict match outcomes. Goddard was able to determine that the best predictions came from a hybrid of the two models; however, the differences in results were small. Pseudo-likelihood statistics were gathered for ten seasons (1992-1993 through 2001-2002). The average pseudo-likelihood statistics were approximately the same for each of the four models in question.

Hirotsu and Wright represented a soccer match as a four-state model which includes a goal scored by team A, possession of the ball for team A, and the same for team B [12]. They constructed a Markov process model, using the four-state model described above, to evaluate the characteristics of teams during a match. By taking into account the number of goals and possession by teams during a match, they were able to evaluate such parameters as offensive and defensive strengths of teams as well as the home team advantage.

3.2.2 Elo Rating

Another popular rating system used to predict the outcome of sports matches is the Elo rating method that was originally developed to rank chess players based on their skill level. Hvattum and Arntzen use the Elo rating system to determine covariates to be used as input to ordered logit regression models [13]. These inputs were the Elo rating differences between the two teams competing in a given match. They constructed an adapted version of the Elo rating system known as goal-based Elo, as well as a basic Elo rating system, which they use to compare to six other benchmark methods. They were able to conclude that the two Elo-based rating systems performed much worse than the two methods based on market odds, yet they performed better than the rest of the methods. They were also able to conclude that the rating difference between two teams is a highly significant predictor of match outcomes.

Leitner et al. constructed a framework to predict the winning probabilities for each of the sixteen team competing in the UEFA EURO 2008 tournament [14]. Using a simulation approach, they were able to construct methods to determine winning probabilities based on a team's abilities (using the Elo rating system) and bookmakers' odds. The model based on bookmakers' odds greatly outperformed the model based on a team's abilities, with the model based on bookmakers' odds correctly predicting

the two teams in the final of the UEFA EURO 2008 tournament.

3.2.3 Poisson Models

A common model used to predict the outcomes of sports matches is the bivariate Poisson model. Adaptations of this model, including the double-Poisson model and diagonal inflated models were used to fit sports data for water polo and soccer. It was determined that the models fit the soccer data to a much greater extent, because they were able to handle overdispersion and correlation [15].

Dixon and Coles also utilized a bivariate Poisson model to predict the outcomes of matches in soccer [16]. Their focus was to use this model to receive positive returns from bets against bookmakers' odds. The bivariate Poisson distribution used goals scored by each team in relation to each teams' performances over three years. The model constructed during this study resulted in positive returns against bookmakers odds over two years.

Other types of predictions can be useful when developing models for soccer matches. Koning et al. developed a simulation/probability model to determine which national team is most likely to win a tournament [17]. Their model assumed a Poisson distribution for the number of goals scored, which is the general consensus among other literature of this nature. The model derived was used to predict the winning probabilities of four major national tournaments in which it was successful three out of four times. This model can also be used to predict the winners of individual matches which is the technique used in this thesis.

3.2.4 Probit Models

Forrest et al. use an ordered probit model to predict the outcomes of English soccer matches [18]. Their model used past match results to predict the result of a future match, and they used this model to compare to the predictions of expert judges.

In order to compare the two prediction methods, they used a series of likelihood-ratio tests in which they were able to conclude that the probabilities output from their ordered probit model used with bookmakers' odds provided the most accurate results.

3.2.5 Factors in Game Prediction

There are many statistics recorded throughout a soccer match. Possession is a significant statistic often used when modeling the outcome of a match. Lago and Martin examine four different variables in order to determine their significance in relation to possession [19]. Those four variables are the home and away teams competing in a match, the location of the match, and the current status of a match. After using linear regression analysis with each of the variables, it was determined that each of four variables was statistically significant in regards to possession of the ball in a given match.

When attempting to develop models to predict outcomes of soccer matches, it is important to know which match statistics reflect winning teams, losing teams, and drawing teams. Casamichana et al. attempted to do just that by analyzing matches between national teams in three different world cups [20]. Their study analyzed two different sets of variables, those related to attacking plays and those related to defensive plays. After performing Levene's test and an analysis of variance, it was determined that the two variables that best represented winning, losing, and drawing teams in relation to attacking play were shots on target and ball possession. Also, the two variables that best represented winning, losing, and drawing teams in relation to defensive play were total shots received and shots on target received.

Home team advantage as it relates to sports is a factor that has been analyzed in great detail over the years. Its relevance in statistical models used to predict the outcomes of sports matches is noted in a study conducted by Beaumont et al. [21].

In combination with attendance data from English Premier League matches, it was determined that home team advantage plays a significant role in the outcome of a match.

Another model was developed by Dixon and Robinson to predict the outcomes of soccer matches [22]. Their model incorporated attacking and defensive parameters for each team, home field advantage, and the current score/time left in the game. They were able to conclude that the rate in which goals are scored increases as the match progresses and is dependent on the current score.

The literature described in this chapter provides numerous crucial concepts which will be analyzed and utilized in predicting the outcomes of soccer matches. Most of the literature agrees that the number of goals scored during a match by each team follows a Poisson distribution based on past matches. Elo ratings, linear regression models, ordered probit models, and Markov process models are all used when predicting the outcomes of soccer matches. Various match statistics used as input to these models are also discussed in the literature.

Chapter 4

Methodology

Two models were implemented for this research. These two models include the Elo Rating Model with multinomial logistic regression (ELO) developed by Hvattum and Arntzen [13], and the Poisson Dixon and Coles (DC) model developed by Dixon and Coles [16]. Both the ELO and DC models were trained and tested on recent EPL data, different from the historic data used in both papers.

4.1 Data

Data for the 2009-2010 through 2017-2018 EPL seasons was obtained from whoscored for this thesis [23]. Figure 4.1 shows a general overview of how this was achieved in this thesis. Each phase of this process is described in more detail below. The data obtained from whoscored.com was separated by each season, with each season being separated by month. Each game contained a variety of statistics that were used in this thesis.

Selenium was chosen as the tool for web scraping because of the various button clicks that needed to be made in order to navigate all seasons and all individual games [24]. In particular, the Python package for Selenium was used. Selenium is a tool which can be used across multiple programming languages to automate web browsing,

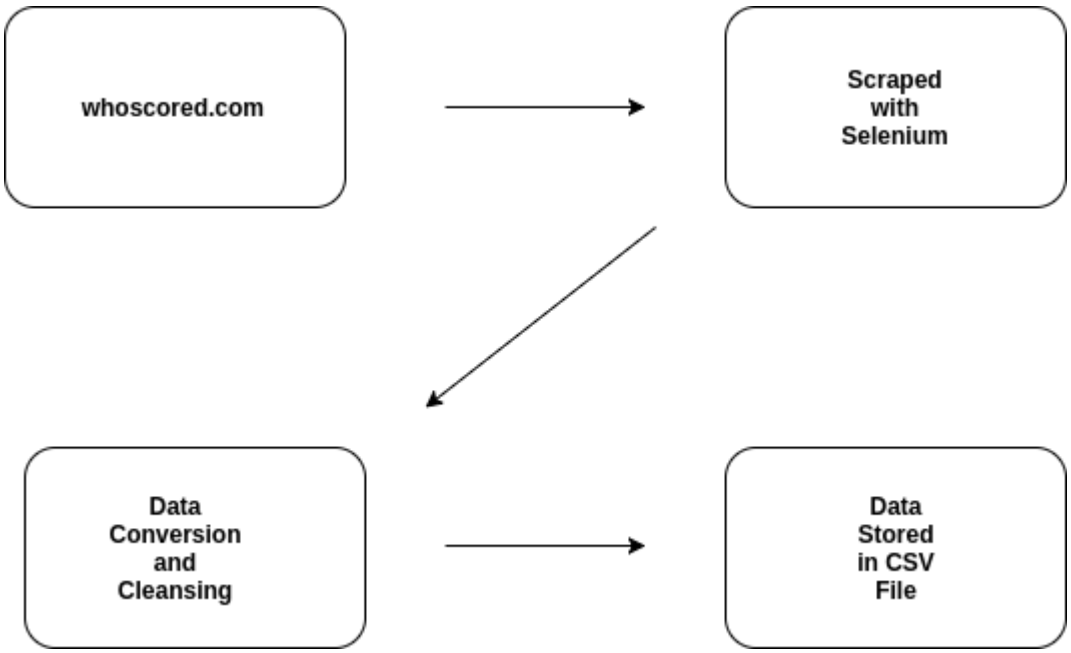

Figure 4.1: Flowchart of Data Acquisition

and it is useful when needed to interact with JavaScript (e.g., button clicks). The web scraper scrapes one season's worth of data at a time, gathering various match statistics from all 380 games played during the EPL regular season. These statistics are then converted into their appropriate data types and appended one game at a time to a Comma Separated Value (CSV) file. One CSV file contains all data for all nine EPL seasons.

date	home_team	away_team	home_team_score	away_team_score
Sat 15-Aug-09	Chelsea	Hull	2	1
Sat 15-Aug-09	Stoke	Burnley	2	0
Sat 15-Aug-09	Portsmouth	Fulham	0	1
Sat 15-Aug-09	Bolton	Sunderland	0	1
Sat 15-Aug-09	Blackburn	Man City	0	2
Sat 15-Aug-09	Wolves	West Ham	0	2
Sat 15-Aug-09	Aston Villa	Wigan	0	2
Sat 15-Aug-09	Everton	Arsenal	1	6
Sun 16-Aug-09	Man Utd	Birmingham	1	0
Sun 16-Aug-09	Tottenham	Liverpool	2	1
Tue 18-Aug-09	Wigan	Wolves	0	1
Tue 18-Aug-09	Sunderland	Chelsea	1	3

Figure 4.2: A Sample CSV File Containing EPL Match Data

The format of the CSV file is shown in Figure 4.2. Data such as the date of the match, the home team, away team, home team score, and away team score are stored in the CSV file, with each row representing one game. Various other match statistics are also stored as extra columns for each match such as shots on target for home and away teams, possession for home and away teams, and number of corner kicks for home and away teams. Sixty-seven unique match statistics were scraped and stored for the purpose of this thesis.

4.2 Elo and Logistic Regression Model

Hvattum and Arntzen describe two ELO methods, one in which goal difference is taken into account and one in which goal difference is not taken into account [13]. For the purpose of this research, the goal-based ELO model was implemented and tested.

4.2.1 Data

The data used for the ELO method includes the home team's name, the away team's name, the home team's score, and the away team's score. This data is split into the following three groups: a training set, an estimate set, and a testing set. Five seasons of data (2009-2010, 2010-2011, 2012-2013, 2014-2015, and 2016-2017) was included in the training set, two seasons of data (2013-2014 and 2015-2016) was included in the estimate set, and two seasons of data (2011-2012 and 2017-2018) was included in the testing set. For the purposes of testing, an additional column was added to the CSV file. This column classified each game as a home win, represented by a "2", a draw, represented by a "1", or an away win, represented as a "0". This was needed in order to test the accuracy of the ELO method after the data was trained and the parameters of the multinomial logistic regression formula were estimated.

4.2.2 The ELO Formula

Each team in the EPL is assigned an ELO rating. The ELO rating of a team reflects the strength of that team based on a set of previous matches. In this thesis, the set of previous matches are included in the training set. ELO ratings are updated for teams after each match played using formula (4.1) from [13],

$$l_1^H = l_0^H + k(\alpha^H - \gamma^H), \tag{4.1}$$

where l_1^Hrepresents the new rating for the home team, l_0^Hrepresents the rating of the home team before the match, k represents the rating update coefficient, α^Hrepresents the actual score of the home team defined in formula (4.2) from [13],

$$\alpha^H = \begin{cases} 1 & \text{if home team won} \\ 0.5 & \text{if draw} \\ 0 & \text{if home team lost} \end{cases} \tag{4.2}$$

γ^H represents the expected score of the home team defined in formula (4.3) from [13],

$$\gamma^H = \frac{1}{1 + c^{(l_0^A - (l_0^H + h))/d}} \tag{4.3}$$

where c and d are both constants, h represents the additional points awarded for home field advantage [13]. In the goal-based ELO model used in this research, k takes into account the score difference, giving more weight to winning and losing by a larger margin as defined in formula (4.4) from [13],

$$k = k_0(1 + \delta)^\lambda \tag{4.4}$$

where k_0 and λ are fixed parameters both greater than zero, and δ represents the absolute goal difference in a match. In order to obtain the same calculations for

the away team, formula (4.5), formula (4.6), and formula (4.7) are used from [13]. Formula (4.5) displays how to calculate the expected score for the away team.

$$\gamma^A = 1 - \gamma^H \tag{4.5}$$

Formula (4.6) displays how to calculate the actual score for the away team.

$$\alpha^A = 1 - \alpha^H \tag{4.6}$$

Formula (4.7) displays how to update the away team's ELO rating after each match.

$$l_1^A = l_0^A + k(\alpha^A - \gamma^A) \tag{4.7}$$

In this research, these parameters are defined as $c = 10$, $d = 400$, and $k_0 = 10$ as described in reference [13].

4.2.3 ELO Implementation

This thesis considers home-field advantage by awarding the home team an additional number of points before calculating their expected number of goals in a match. In order to calculate the proper number of home advantage points for a given match, the percentage of home wins and the percentage of away wins was gathered from the training data. The percentage of draws was split evenly between the home win percentage and the away win percentage. Then, using (4.3) with $l_0^H = 1500$ and $l_0^A = 1500$, h was set to one and continuously incremented until the expected number of goals equaled the home win percentage calculated from the training data. The resulting value was used as the home advantage parameter.

Figure 4.3 displays an overview of how the ELO model was implemented. The home advantage calculation just described is used to obtain initial ELO calculations

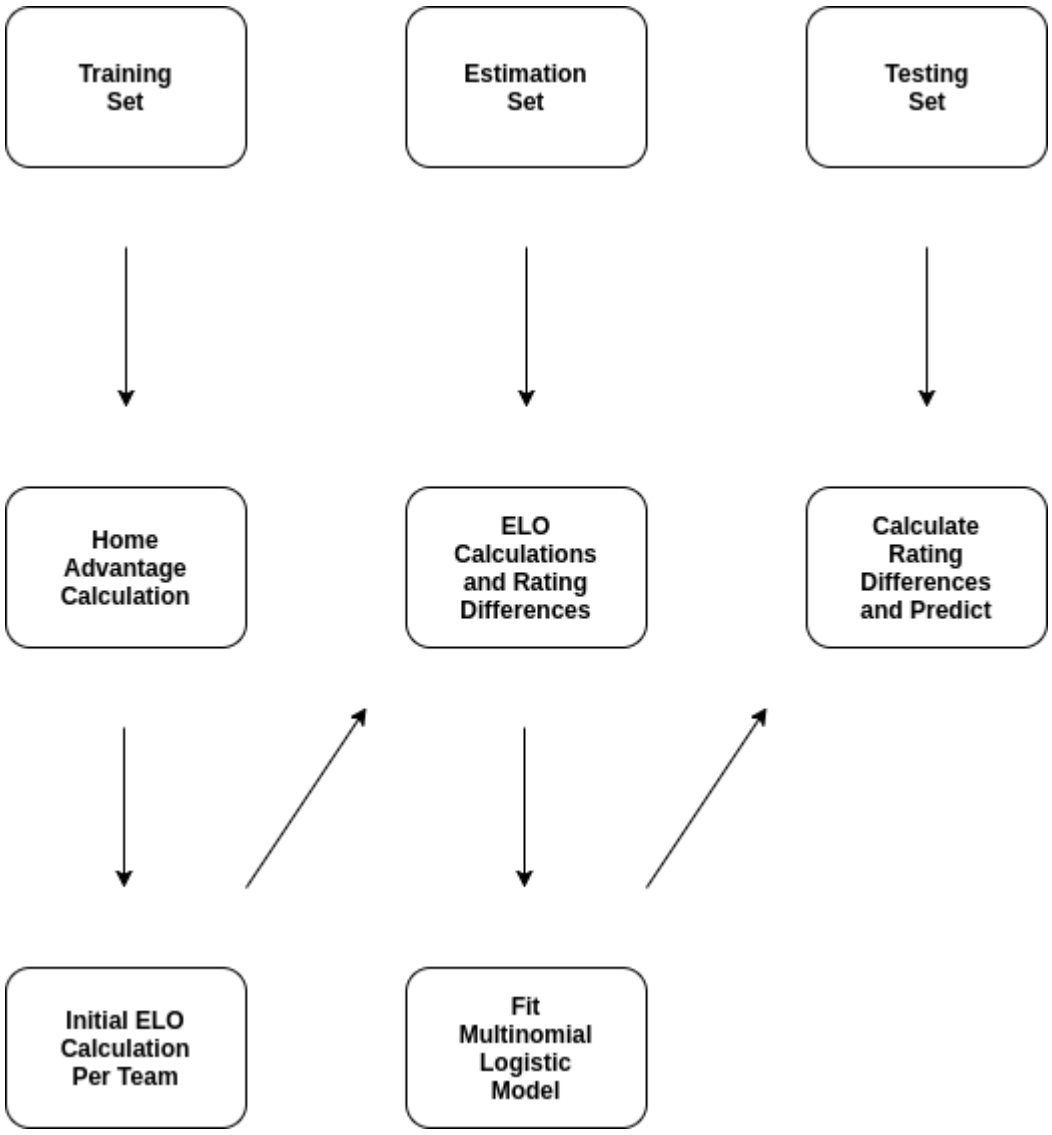

Figure 4.3: ELO Method Flowchart

for each team based on the set of training matches totaling 1140 games. Teams in the training set that are not in the testing set are given a base 1500 points before predictions are made. Next, using the initial ELO calculations obtained from the training set, the parameters for the multinomial logistic regression model are estimated with data from the estimation set. The only covariate used in this estimation is the rating difference of the two teams competing in a match in favor of the home team, x, defined in formula (4.8).

$$x = l_0^H - l_0^A \tag{4.8}$$

The ELO ratings for teams were updated after each match in addition to compiling a list of rating differences. The list of rating differences and the classification for each match were input into a popular LogisticRegression fit function in order to properly estimate the parameters of the multinomial logistic regression model [25]. This thesis uses multinomial logistic regression instead of ordinal logistic regression, which was used by Hvattum and Arntzen, because the only area of interest was to generate home win, away win, and draw percentages for each EPL match played.

Finally, the matches within the testing set were simulated using the ELO ratings found from the training and estimation sets. Once again, a list of rating differences was compiled and input into the LogisticRegression score function along with the actual match classifications in order to obtain the mean accuracy for the testing set.

4.3 Dixon & Coles Poisson Model (DC)

A basic Poisson model results in a probability distribution for a certain number of events occurring in a fixed interval of time with some known average rate. Dixon and Coles assume that the number of home goals and the number of away goals scored in a 90 minute soccer match follow independent Poisson distributions [16].

4.3.1 The Data for the DC Model

The data used in the DC model is similar to the data used in the ELO model. This includes the home team's name, the away team's name, the home team's score, and the away team's score. The seven seasons worth of data is split into a training set comprised of five seasons (1900 games), and a testing set comprised of two seasons (760 games). An additional categorical column was not needed in the CSV file for

the DC model, because no form of logistic regression was used.

4.3.2 DC Model Implementation

The DC model assumes that the average number of home and away goals follow independent poisson distributions. The probability of the home team scoring a certain number of goals and the away team scoring a certain number of goals can be calculated using the formula described in formula (4.9) from [16],

$$Pr(X_{i,j} = x, Y_{i,j} = y) = \tau_{\lambda,\mu}(x, y)\frac{\lambda^x e^{-\lambda}}{x!}\frac{\mu^y e^{-\mu}}{y!}, \tag{4.9}$$

where x is the number of goals scored by the home team, y is the number of goals scored by the away team, and λ and μ are defined in formula (4.10) and formula (4.11),

$$\lambda = e^{\alpha_i+\beta_j+\gamma} \tag{4.10}$$

$$\mu = e^{\alpha_j+\beta_i} \tag{4.11}$$

where i and j represent the home and away teams respectively, α and β represent the attack and defense ratings of teams respectively, and γ represents the home field advantage parameter.

4.3.3 DC Adjustment

In Equation 4.9, τ represents the adjustment for low scoring games found by Dixon and Coles [16]. Dixon and Coles found that their assumption of independence for the number of home and away goals scored in a game was accurate except for games ending in 0-0, 0-1, 1-0, and 1-1, thus τ represents the following adjustment in formula (4.12) from [16]

$$\tau_{\lambda,\mu}(x,y) = \begin{cases} 1-\lambda\mu\rho & \text{if x = y = 0} \\ 1+\lambda\rho & \text{if x = 0, y = 1} \\ 1+\mu\rho & \text{if x = 1, y = 0} \\ 1-\rho & \text{if x = y = 1} \\ 1 & \text{otherwise} \end{cases} \tag{4.12}$$

where ρ represents a dependence parameter and $\rho = 0$ represents independence [16]. Using this adjustment, this model is able to improve the accuracy of predicting the low scoring match outcomes mentioned above.

4.3.4 The Model

Each team in the EPL for the seasons contained in the training and testing sets receive an attack and defense rating. As a result, the attack and defense ratings for each team, as well as the dependence parameter ρ and the home field advantage parameter γ need to be estimated in order to make predictions for the outcomes of soccer matches. These parameters can be estimated using the four seasons of training data to maximize the following equation from [16],

$$L(\alpha_i, \beta_i, \rho, \gamma; i = 1, ..., n) = \prod_{k=1}^{N} \tau_{\lambda_k \mu_k}(x_k, y_k) e^{-\lambda_k} \lambda_k^{x_k} e^{-\mu_k} \mu_k^{y_k}, \tag{4.13}$$

where N is the total number of games played, n is the total number of teams, k is the index of the match being played, and λ_k and μ_k are described in formula (4.14) and formula (4.15),

$$\lambda_k = e^{\alpha_{i(k)} + \beta_{j(k)} + \gamma} \tag{4.14}$$

$$\mu_k = e^{\alpha_{j(k)}+\beta_{i(k)}} \tag{4.15}$$

where $i(k)$ and $j(k)$ represent the home and away teams playing in match k. In order to ensure that the model is not overparameterized, a constraint is added such that 1/sum of all team's attack ratings is 1.

In order to properly estimate the parameters specified above, all of the data from the CSV file is loaded into a pandas dataframe. Pandas is an open-source Python library used in data analysis. The pandas dataframe stores the data in a two dimensional data structure and allows columns to be referenced by headings. This makes it simple to extract and manipulate the data contained within the dataframe.

Figure 4.4 describes the process of implementing the DC model. Initial estimates for all parameters need to be set in order to properly estimate the attack and defense parameters for each team, the dependence parameter, and the home field advantage parameter. For this research, all attack and defense parameters were initialized to 0.0, the dependence parameter was initialized to 0.0, and the home field advantage parameter was initialized to 0.0. A constraint was added to the dependence parameter such that it could not be greater than 1 or less than negative one. All initial parameters were combined to create one list of the form described in Figure 4.5.

Next, two matrices were constructed, one containing home team data and another containing away team data. In both matrices, each column represents a different team, and each row represents an individual game. If that particular team was active in that game, a 1 is placed in their column in the home team matrix if they were the home team in the game, or a 1 is placed in their column in the away team matrix if they were the away team in the game. This data is then concatenated with a list containing home team goals and away team goals from all games as well as all team names. This list takes the form described in Figure 4.6

Then, using SciPy's *optimize.minimize* function, the parameters in Figure 4.5 are

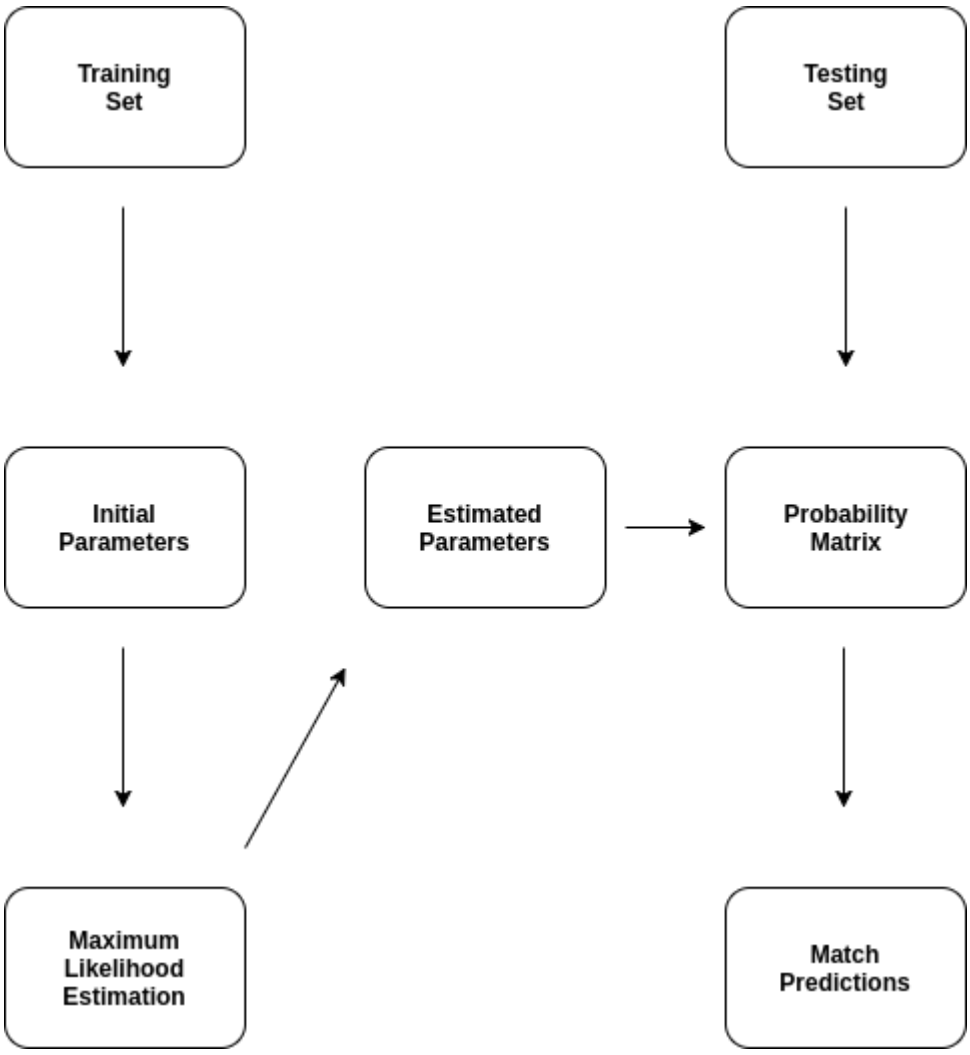

Figure 4.4: DC Method Flowchart

$$[\gamma, \rho, \alpha_1, \alpha_2, ..., \alpha_n, \beta_1, \beta_2, ..., \beta_n]$$

Figure 4.5: The List of Parameters

[*home team matrix, away team matrix, home goals, away goals, team names*]

Figure 4.6: The Data List

estimated based on the data from the list in Figure 4.6 [26]. This function takes in an optimizing function, the parameters to be estimated, the data to be used in the parameter estimation, and any constraints that need to be considered during optimization. When maximizing the likelihood function described in Equation 4.13, it is simpler to work with the log-likelihood. The log-likelihood equivalent of Equation

4.13 is shown in formula (4.16).

$$log(L) = \sum_{k=1}^{N} log(\tau_{\lambda_k \mu_k}(x_k, y_k)) + (-\lambda_k log(e)) + (x_k log(\lambda_k)) + (-\mu_k log(e)) + (y_k log(\mu_k)) \tag{4.16}$$

The *optimize.minimize* function then uses the log-likelihood function specified in Equation 4.16 to obtain estimates for each of the parameters. Using these estimates, probability matrices can be constructed for each match in the testing set. This matrix is of the form described in Figure 4.7.

$$\begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{array} \begin{array}{c} \begin{array}{ccccccc} 0 & 1 & 2 & 3 & 4 & 5 & 6 \end{array} \\ \begin{pmatrix} d & a & a & a & a & a & a \\ h & d & a & a & a & a & a \\ h & h & d & a & a & a & a \\ h & h & h & d & a & a & a \\ h & h & h & h & d & a & a \\ h & h & h & h & h & d & a \\ h & h & h & h & h & h & d \end{pmatrix} \end{array}$$

Figure 4.7: Example Probability Matrix

The columns in Figure 4.7 represent the number of away goals scored and the rows represent the number of home goals scored. The intersection of each row and column represents the probability of the outcome occurring. The probability of an away win, draw, and home win can be determined by adding up all of the as, ds, and hs, respectively.

Chapter 5

Results

Results were obtained for both models implemented in this research. For each model, CSV files were generated containing predictions from each game in the testing dataset. A few observations are also made about games resulting in draws, because neither of the two implemented models predicted draws with a high accuracy.

5.1 Elo & Logistic Regression Model

Upon completion of model, results were stored in a CSV file, a few of which can be seen in Table 5.1. The home and away teams, the probability of a home win, away win, and a draw, and the actual result of the game were all were all included in the CSV file.

Table 5.1: A Sample of Elo Multinomial Logistic Regression Game Predictions

Home Team	Away Team	Home Win	Away Win	Draw	Actual
Arsenal	Leicester	45.43%	28.56%	26.00%	Home Win
Stoke	Southampton	27.44%	44.89%	27.67%	Home Win
Liverpool	Man Utd	39.54%	33.47%	26.99%	Draw
West Brom	Man City	14.17%	60.60%	25.24%	Away win
Watford	Stoke	47.95%	26.58%	25.47%	Away Win
Chelsea	Swansea	66.54%	13.75%	19.71%	Home Win
Arsenal	Liverpool	45.49%	28.52%	25.99%	Draw
West Brom	Swansea	34.82%	37.69%	27.49%	Draw
Leicester	Burnley	63.60%	15.57%	20.82%	Home Win
Crystal Palace	Watford	37.01%	35.70%	27.29%	Home Win

The probability column in Table 5.1 that has the highest percentage was used as the predicted result for each game in the testing dataset. It is important to note that newly promoted teams with no previous Elo rating received the same initial rating of 1500 just as every other team did. If a team was in the EPL at any point in games within the testing dataset, but not in any games within the training dataset, they still received an initial rating of 1500.

The final Elo ratings for teams over the course of the 2011-2012 and 2017-2018 EPL seasons are shown in Table 5.2.

Table 5.2: Final Elo Ratings 2011-2012 2017-2018 EPL Seasons

Team	Rating	Team	Rating
Man City	2088	Man Utd	1948
Tottenham	1926	Liverpool	1899
Arsenal	1823	Chelsea	1803
Leicester City	1731	Everton	1719
Crystal Palace	1695	Newcastle	1695
Southampton	1688	West Ham	1680
Bournemouth	1646	Burnley	1643
Sunderland	1623	Fulham	1599
Swansea	1598	Watford	1597
Brighton	1595	West Brom	1583
Stoke City	1579	Wigan Athletic	1559
Norwich City	1556	Huddersfield	1553
Aston Villa	1513	Bolton	1501
Blackburn	1490	Queens Park Rangers	1487
Wolves	1408	-	-

An additional 67 points were awarded to the home team in each game to account for home-field advantage. In other words, on average, the percentage of home wins and the percentage of away wins were equal when the home team had 67 more rating points than the away team. The overall accuracy of this model when predicting a win, loss, or draw was 53.03%. The key values obtained from running this model on the testing dataset are better shown in Table 5.3.

Table 5.3: Estimated Parameters and Model Accuracy for Elo Multinomial Logistic Regression Model

Parameter	Value
Home Advantage	67.00
Model Accuracy	53.03%
Home Win Accuracy	73.84%
Away Win Accuracy	58.04%
Draw Accuracy	0.00%

Also shown in Table 5.3 are the specific accuracies for home win, away wins, and draws. In other words, of the games that actually resulted in a home win, the model was able to predict those 73.84% of the time. The model was also able to predict 58.04% of the actual away wins, and 0% of the draws.

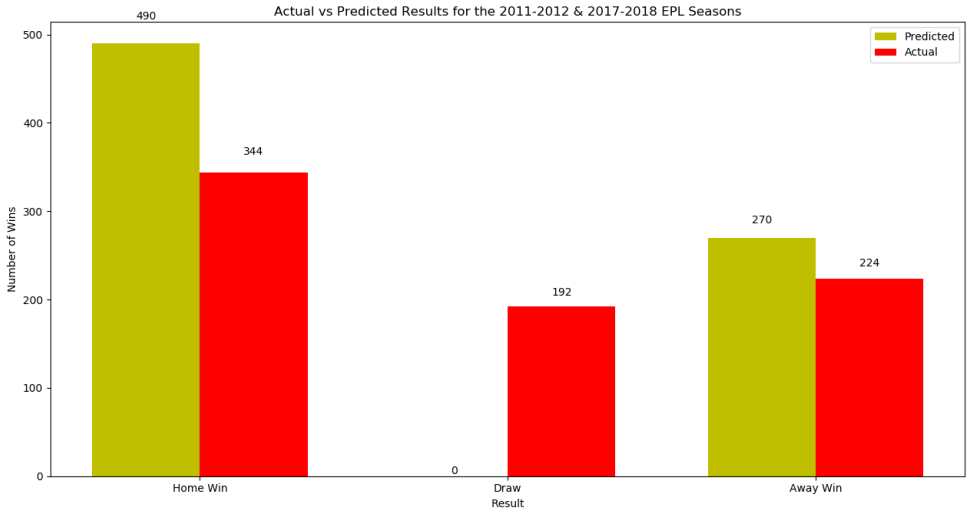

Figure 5.1: Predicted vs Actual Results Elo

The predicted and actual results for this model on the testing data are shown in Figure 5.1. The number of away wins predicted are slightly overestimated, but

relatively close when compared to the greatly over predicted number of home wins. The model did not predict any draws.

5.2 Dixon & Coles Poisson Model

The results for the Dixon and Coles Poisson model are displayed in Table 5.4. Data was stored similarly to the Elo model results in a CSV file with home and away teams, match outcome probabilities, and the actual result of the game.

Table 5.4: Dixon and Coles Poisson Model Game Predictions

Home Team	Away Team	Home Win	Away Win	Draw	Actual
Arsenal	Leicester	70.19%	10.59%	17.10%	Home Win
Stoke	Southampton	36.09%	31.70%	32.18%	Home win
Liverpool	Man Utd	35.49%	36.44%	28.00%	Draw
West Brom	Man City	22.59%	51.68%	25.30%	Away Win
Watford	Stoke	34.84%	34.97%	30.15%	Away Win
Chelsea	Swansea	74.05%	7.87%	15.53%	Home Win
Arsenal	Liverpool	52.48%	22.66%	24.36%	Draw
West Brom	Swansea	45.04%	27.67%	27.09%	Draw
Leicester	Burnley	54.86%	19.56%	25.03%	Home Win
Crystal Palace	Watford	55.16%	20.22%	25.57%	Home Win

From Table 5.4, the probability column with the highest percentage was used as the predicted result for each game.

Table 5.5 displays the estimated attack and defense strength for each team. However, because both Brighton and Huddersfield were not in the EPL during any season contained in the training dataset, their attack and defense ratings were not previously estimated. Instead, they were assigned random values between 0.1 and 0.9 for attack

strength and between -0.9 and -0.1 for defense strength.

Table 5.5: Attack and Defense Ratings for EPL Teams

Team	Attack	Defense	Team	Attack	Defense
Arsenal	1.44	-1.17	Aston Villa	0.93	-0.87
Birmingham	0.76	-0.93	Blackburn	0.91	-0.84
Blackpool	1.20	0.00	Bolton	0.98	-0.76
Bournemouth	1.20	-0.69	Brighton	0.79	-0.39
Burnley	0.76	-0.75	Chelsea	1.51	-1.34
Crystal Palace	1.02	-0.85	Everton	1.14	-1.06
Fulham	0.96	-0.98	Huddersfield	0.15	-0.28
Hull City	0.72	-0.66	Leicester City	1.02	-0.82
Liverpool	1.29	-1.13	Manchester City	1.40	-1.24
Manchester United	1.41	-1.32	Middlesbrough	0.51	-0.94
Newcastle United	0.99	-0.77	Norwich City	0.91	-0.84
Portsmouth	0.72	-0.68	Queens Park Rangers	0.77	-0.71
Reading	0.97	-0.61	Southampton	1.02	-1.05
Stoke City	0.83	-1.03	Sunderland	0.80	-0.85
Swansea City	0.99	-0.87	Tottenham	1.32	-1.13
Watford	0.89	-0.69	West Bromwich Albion	1.00	-0.85
West Ham United	0.96	-0.80	Wigan Athletic	0.87	-0.64
Wolves	0.80	-0.79	-	-	-

After optimizing Equation 4.16, it was determined that the estimate for the dependence parameter ρ was -0.07 and the estimate for the home field advantage parameter γ was 0.29. The overall accuracy of this model when predicting the games in the training dataset was 51.97% . All of these estimated parameters are shown in Table 5.6.

Table 5.6: Estimated Parameters and Model Accuracy for Dixon and Coles Poisson Model

Parameter	Value
Dependence (ρ)	-0.07
Home Advantage (γ)	0.29
Model Accuracy	51.32%
Home Win Accuracy	81.10%
Away Win Accuracy	49.55%
Draw Accuracy	0.00%

In Table 5.6, similar to the results gathered for the ELO method, a home win accuracy of 81.10%, an away win accuracy of 49.55%, and a draw accuracy of 0.00% was recorded.

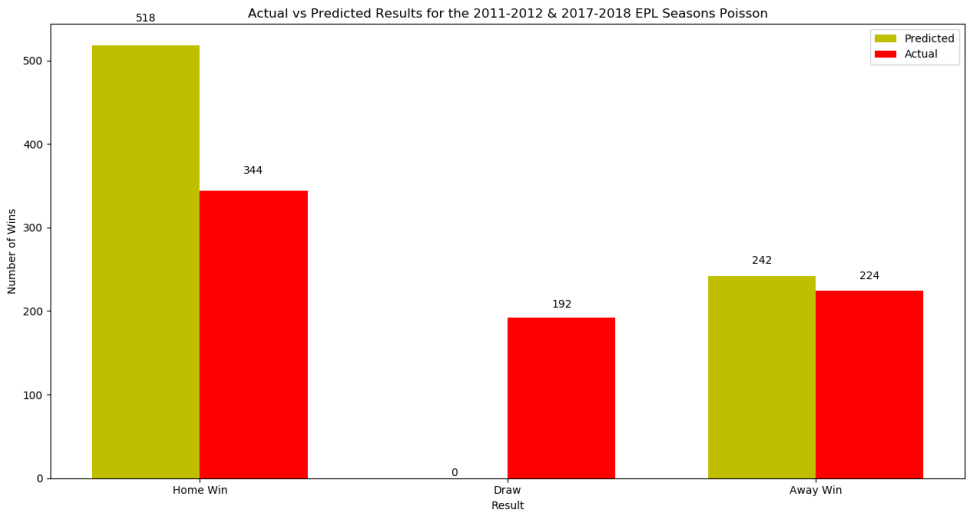

Figure 5.2: Predicted vs Actual Results Poisson

A visual of the final predictions for the 2011-2012 and 2017-2018 seasons is shown in Figure 5.2. The model closely predicted the number of away wins, no draws were

predicted, and the number of home wins were greatly over predicted. While this is also the case for the Elo with multinomial logistic regression model, the Dixon and Coles poisson model overpredicted the number of home wins to a greater extent.

Chapter 6

Conclusions

After training, estimating the parameters, and running both the Elo with multinomial logistic regression and Dixon and Coles Poisson models, similar results were obtained. Neither model was able to predict the occurrence of a draw. While both models greatly overestimated the number of home wins during the 2011-2012 and 2017-2018 EPL seasons, the Dixon and Coles poisson model did so to a greater extent. For both models, games that ended in draws tended to be predicted as home wins. This could be due to an overestimated home field advantage parameter in both models.

Both models achieve a similar level of accuracy in terms of predicting home wins, away wins, and draws as shown in Tables 5.4 and 5.8 with the Elo model performing marginally better. While the difference between the two models in terms of accuracy is minuscule, the slightly better performance of the Elo model may be attributed into margin of victory weighing more heavily on the impact of team rating fluctuation.

The Elo ratings from Table 5.2 as well as the attack and defense ratings for each team from Table 5.5 were accurate predictors for determining which teams would finish in the top four in the league table for the 2011-2012 and 2017-2018 EPL seasons. The stronger teams throughout the years (Arsenal, Chelsea, Liverpool, Tottenham, Manchester United, and Manchester City) all finished with the highest Elo ratings as

well as the greater attack and defense ratings.

Other similarities between the Elo with multinomial logistic regression model and the Dixon and Coles poisson model are shown in Tables 5.2, 5.3, 5.5, and 5.6. The final Elo ratings and attack/defense strengths among top teams such as Manchester United, Manchester City, Liverpool, and Chelsea are consistant across both models. These ratings also accurately reflect their average placings across the 2011-2012 and 2017-2018 EPL seasons. The same is true of teams finishing near the bottom of the table in both seasons such as Brighton, West Brom, and Wolves. This results in being an accurate predictor for determining which teams are likely to be relegated at the end of the season (the bottom three teams in the standings), and which teams will earn qualification for the Champions League (the top four teams in the standings).

6.1 Future Work

Both models can be improved by introducing other key in-game statistics such as time of possession, number of shots on goal, number of corners, etc. Determining significant factors outside of home field advantage and margin of victory would have potential to aid in the overall accuracy of both models.

Another area of improvement with both models would be to consider various time constraints. With the current implementation of the models, the testing, estimation, and training datasets contain non-concurrent seasons. This is most prominent in the Dixon and Coles poisson model where the attack and defense ratings of teams are estimated over the games in the training dataset, but they do not change over the course of the games within the testing dataset. Re-estimating the attack and defense ratings after each game would have the potential to improve the accuracy of the model.

Developing a more sophisticated system for assigning initial Elo ratings in the Elo

model, and initial attack and defense ratings in the Dixon and Coles poisson model would likely aid in the overall accuracy of both models as well. Teams are relagated from and promoted to the EPL every season. These teams were assigned an initial rating at 1500 in the Elo model, and a random value for attack and defense ratings in the Dixon and Coles poisson model. A more sophisticated system may take the ratings of the relegated teams and assign them to the newly promoted teams.

Another area that can be expanded upon would be the use of ordered logistic regression in place of multinomial logistic regression in the ELO model. By doing so, the two approaches could be compared in order to determine which is the more logical to use in regards to predicting the outcomes of soccer matches.

Bibliography

[1] SAS, "Predictive Analytics: What it is and why it matters," https://www.sas.com/en_us/insights/analytics/predictive-analytics.html, Oct. 2018.

[2] FiveThirtyEight, "How Our College Football Playoff Predictions Work," https://fivethirtyeight.com/methodology/how-our-college-football-playoff-predictions-work/, Oct. 2018.

[3] I. MATHEMATICS, "The Poisson Probability Distribution," https://www.intmath.com/counting-probability/13-poisson-probability-distribution.php, Jan. 2018.

[4] D. Barrow, I. Drayer, P. Elliott, G. Gaut, and B. Osting, "Ranking rankings: an empirical comparison of the predictive power of sports ranking methods," *Journal of Quantitative Analysis in Sports*, vol. 9, Jan. 2013.

[5] R. Stefani and R. Pollard, "Football Rating Systems for Top-Level Competition: A Critical Survey," *Journal of Quantitative Analysis in Sports*, vol. 3, Jan. 2007.

[6] P. S. Gill, "Late-Game Reversals in Professional Basketball, Football, and Hockey," *The American Statistician*, vol. 54, pp. 94–99, May 2000.

[7] J. H. Lebovic and L. Sigelman, "The forecasting accuracy and determinants of football rankings," *International Journal of Forecasting*, vol. 17, pp. 105–120, 2001.

[8] C. Song, B. L. Boulier, and H. O. Stekler, "The comparative accuracy of judgmental and model forecasts of American football games," *International Journal of Forecasting*, vol. 23, pp. 405–413, Jul. 2007.

[9] B. L. Boulier and H. Stekler, "Predicting the outcomes of National Football League games," *International Journal of Forecasting*, vol. 19, pp. 257–270, Apr. 2003.

[10] C. Song, B. L. Boulier, and H. O. Stekler, "Measuring consensus in binary forecasts: NFL game predictions," *International Journal of Forecasting*, vol. 25, pp. 182–191, Jan. 2009.

[11] J. Goddard, "Regression models for forecasting goals and match results in association football," *International Journal of Forecasting*, vol. 21, pp. 331–340, Apr. 2005.

[12] N. Hirotsu and M. Wright, "An evaluation of characteristics of teams in association football by using a Markov process model," *Journal of the Royal Statistical Society: Series D (The Statistician)*, vol. 52, pp. 591–602, Dec. 2003.

[13] L. M. Hvattum and H. Arntzen, "Using ELO ratings for match result prediction in association football," *International Journal of Forecasting*, vol. 26, pp. 460–470, Jul. 2010.

[14] C. Leitner, A. Zeileis, and K. Hornik, "Forecasting sports tournaments by ratings of (prob)abilities: A comparison for the EURO 2008," *International Journal of Forecasting*, vol. 26, pp. 471–481, Jul. 2010.

[15] D. Karlis and I. Ntzoufras, "Analysis of sports data by using bivariate Poisson models," *Journal of the Royal Statistical Society: Series D (The Statistician)*, vol. 52, pp. 381–393, Oct. 2003.

[16] M. J. Dixon and S. G. Coles, "Modelling Association Football Scores and Inefficiencies in the Football Betting Market," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, vol. 46, pp. 265–280, Jan. 1997.

[17] R. H. Koning, M. Koolhaas, G. Renes, and G. Ridder, "A simulation model for football championships," *European Journal of Operational Research*, vol. 148, pp. 268–276, Jul. 2003.

[18] D. Forrest, J. Goddard, and R. Simmons, "Odds-setters as forecasters: The case of English football," *International Journal of Forecasting*, vol. 21, pp. 551–564, Jul. 2005.

[19] C. Lago and R. Martn, "Determinants of possession of the ball in soccer," *Journal of Sports Sciences*, vol. 25, pp. 969–974, Jul. 2007.

[20] J. Castellano, D. Casamichana, and C. Lago, "The Use of Match Statistics that Discriminate Between Successful and Unsuccessful Soccer Teams," *Journal of Human Kinetics*, vol. 31, Jan. 2012.

[21] D. Forrest, J. Beaumont, J. Goddard, and R. Simmons, "Home advantage and the debate about competitive balance in professional sports leagues," *Journal of Sports Sciences*, vol. 23, pp. 439–445, Apr. 2005.

[22] M. Dixon and M. Robinson, "A birth process model for association football matches," *Journal of the Royal Statistical Society: Series D (The Statistician)*, vol. 47, pp. 523–538, Sep. 1998.

[23] WhoScored, "Football Statistics | Football Live Scores | WhoScored.com," https://www.whoscored.com/, Aug. 2018.

[24] Selenium-Python, "Selenium with Python Selenium Python Bindings 2 documentation," http://selenium-python.readthedocs.io/, Aug. 2018.

[25] Sklearn, "sklearn.linear_model.LogisticRegression scikit-learn 0.18.2 documentation," http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html#sklearn.linear_model.LogisticRegression.predict_proba, Oct. 2018.

[26] SciPy, "scipy.optimize.minimize scipy v1.1.0 reference guide," https://docs.scipy.org/doc/scipy/reference/generated/scipy.optimize.minimize.html, Nov. 2018.

Vita

Zachary Andrews was born in 1993, in Raleigh, North Carolina to Ken and Debi Andrews. He was accepted to Appalachian State University in 2012 where he began his Bachelor of Science in Computer Science. He earned his Bachelor of Science in Computer Science in 2015 and immediately proceeded into graduate studies at Appalachian State University, graduating in 2019.