University of Tartu

Faculty of Science and Technology

Institute of Mathematics and Statistics

Farhad Tahirov

Application of Poisson and Dixon-Coles models on football match outcome prediction and research of a positive return over investment in betting market

Actuarial and Financial Engineering

Master's Thesis(30 ECTS)

Supervisor Jüri Lember

Tartu 2020

ACKNOWLEDGEMENT

This project would not have been possible without the help and cooperation of many. I would like to thank the people who helped me directly and indirectly in the completion of this project work.

First and foremost, I would like to express my gratitude to my supervisor, **Dr. Jüri Lember**, for his constant support and help throughout the duration of the project. His confidence in my work encouraged me to do my best.

Secondly, I would like to thank our program coordinator, Meelis Kaarik, for his expertise and for taking the time to discuss my ideas and to give me interesting insights.

I would also like to thank all the faculty members of the İnstitute of Mathematics and Statistics and my classmates for their steadfast and strong support and engagement with this project.

Abstract

Data analysis has become the main driver of successful decision making in our nowadays world. From startups to big businesses application of statistics over constantly accumulating data has proven to be the key for growth in many industries. Currently, alongside business organizations and high-tech firms, governmental institutions, medical industry and many more rely on insights derived from big data. Usage of proper statistical models over data can increase a firm's profitability, identify a medical test's accuracy, support banks recognize fraud transactions and many more.

One of the platforms where application of data analysis has grabbed a great deal of attention is over the most popular sport on earth-Football. Application of statistical models in order to predict football match results has been the center of attention for many people, from top scientists to bookmakers already for quite some time. Certain techniques have been proposed to find potential statistical models that could be helpful in predicting match score outcomes. And with growing betting industry many have tried to beat bookies with the help of statistical models developed for making prediction for match results.

In this paper, indirect approaches, namely Poisson and Dixon-Coles models will be applied to predict match score results. The reason why those models are referred as indirect is due to the fact that regression outputs through those models are goals, rather than direct match outcomes. We will try to beat punctuality of decisions derived from one's "gut feeling", an ambiguous term we will formalize in this paper, through using indirect approaches for match outcome modelling. And at the end, it is found that betting strategy formulated with the use of predictions through such models can yield a positive return through betting in the Premier League over the season 2018-2019.

CERCS research specialisation: P160 Statistics, operations research, programming, actuarial mathematics.

Keywords : Data Analysis, Football

Poissoni ja Dixon-Cole mudelite kasutamine jalgpallitulemuste ennustamisel ja spordikihlvedude tulemuslike investeerimistrateegiate uurimine

Magistritöö

Farhad Tahirov

Lühikokkuvõte. Tänapäeva maailmas on andmeanalüüsist saanud edukate otsuste tegemise peamine mootor. Mitmes valdkonnas on just kasvavate andmehulkade statistiline analüüs edu võti, seda nii idufirmades kui ka suurettevõtetes.

Tänapäeval toetuvad suurandmetest saadud teadmistele lisaks äriorganisatsioonidele ja kõrgtehnoloogiaettevõtetele ka valitsusasutused, meditsiinitööstus ja paljud teised valdkonnad. Korralike statistiliste mudelite kasutamine võib näiteks suurendada ettevõtte kasumlikkust, tuvastada meditsiinilise testi täpsust ja aidata pankadel tuvastada pettusi, näiteid on teisigi.

Andmeanalüüsi rakendamine on pälvinud suurt tähelepanu ka kõige populaarsema spordiala – jalgpalli - maailmas. Statistiliste mudelite rakendamine jalgpallimatši tulemuste ennustamisel on juba pikka aega paljude inimeste, tippteadlastest kihlvedude vahendajateni, fookuses. Jalgpallimatši tulemuste ennustamiseks on välja pakutud palju erinevaid mudeleid ja tehnikaid ning spordikihlvedude turu kasvu tõttu proovib nende mudelite abil tulu saada üha rohkem ja rohkem inimesi Käesolevas magistritöös kasutatakse kihlvedude võitmiseks kaudseid lähenemisviise, nimelt Poissoni ja Dixon-Colesi mudeleid. Neid mudeleid nimetatakse kaudseteks seetõttu, et nende abil modelleritakse täpset jalgpallimatši tulemust, mitte aga ainult võitjat või kaotajat. Nimetatud kaudsete mudelite abil üritame saavutada suuremat ennustustäpsust kui see on nö kõhutunde“ meetodidel, millised formuleerime töö käigus. Magistritöö lõpus näitame Inglise kõrgliiga 2018-2019 hooaja tulemuste kaudu, et nimetatud kaudsed mudelid võivad kihla vedades anda reaalset tulu.

CERCS uurimisvaldkonnad: Statistika, operatsioonianalüüs, programmeerimine, kindlustusmatemaatika

Märksõnad: andmeanalüüs, jalgpall

:

TABLE OF CONTENTS

Chapter 1

Introduction

1.1 Statistics for Football

Being the most popular sport on earth, football has gathered quite some attention especially in the 21st century. Recently, various types of data such as player statistics, number of shots and etc. have been gathered by bookmakers, statisticians. Consequently, availability of such enormous information have allowed data scientists to predict outcomes as such:

- Match Strategy
- Player Performance
- Corner Predictions
- Above/Below 2.5 goals predictions
- Match outcome prediction
- Exact score prediction
- Betting odds calculation

Currently, betting market is worth billions of dollars and seems to be growing over time. Additional availability of access for betting on live matches have further paved the way for the market to attract interest.

1.2 Motivation

An important element of application of statistics on football matches is to attempt deciphering information regarding the match outcomes based on historical data. This process is referred to as predicting future match outcomes relying on historical data and statistical

models applied on the data.

It is indeed understandable that there is a big random element in football matches, and statistical models cannot perfectly control those random elements which occur in real life. That being said, it could be quite challenging to predict match outcomes in the presence of so-called surprise events. Such events can be described as a key player being injured, a player receiving a red card, weak team gaining luck all of a sudden and etc. As famous saying goes: "The ball is round and the game lasts 90 minutes". In other words, no matter how large data we have, random element is always there and it is quite of an obstacle to account for that randomness in statistical models. However, possessing randomness does not make football matches be impossible to predict, at the very least, certain accuracy level can be expected to be reached and a potential investment strategy through betting can be researched.

At the first glance, it might be a bit tricky to understand the variable being predicted by the statistical model. Using indirect approaches such as Poisson models, the variable being predicted is the expected number of goals each team will score against the rest of the teams. Inserting these expected number of goals into Poisson Probability Mass Function of goals(for both home and away team), one can extract probabilities of number of goals each team can score. Through using historical matches, we find an optimal method that assigns probabilities for each potential match score in the season 2018-2019 in the Premier League. And then, using these probabilities one can extract probabilities for win, draw and loss result.

1.3 Objectives

This project aims to extend the state of the art by applying indirect methods to model match outcomes using historical data where matches are recorded in seasons 2010-2017 in English Premier League and then make predictions on 2018-2019 season.

The main objective of this paper is to research application of Poisson and Dixon-Coles models to predict the outcome of a football game and observe whether such models beat the benchmark model. In doing so, we initially will create simple models, for example

a model that will always assign win result for home team, or a model that will simply make decisions based on team bookmaker's odds, etc. and such models will be referred as simple models serving as the benchmark. One of the most famous indirect approaches for match outcome prediction is Poisson model. In Poisson model(s) we have developed attacking strength and defensive strength metrics, which are crucial in finding the probability for potential match outcomes. Although very useful, Poisson model inherits certain drawbacks in modelling direct outcome of the game. For example, implicit extraction of match outcomes from probability distribution of number of goals each team can score in a game proves to be underestimating the occurrence of Draw results. Thus, by relying on modifications developed by Mark J.Dixon and Stuart G.Coles, we will adjust Basic Poisson model, and build a modified version of Poisson model, so-called Dixon-Coles model.

It should certainly be noted that when building above-mentioned models, we will be referring to a dataset that includes information only regarding home team, away team and historical match results. As a next step, we will research the possibility of reaching a positive return over investing in betting market with a strategy purely followed by the model results.

1.4 Challanges

• Optimal Betting Strategy: The results the models will produce might attract one to try and see if those models can be proven to yield in a positive return once a proper investment strategy is chosen, relying on those models. Although there might be an optimal investment strategy once predicted probabilities for match results are extracted, at this point we will not focus on optimizing our investment strategy on the basis of probabilities to reach the best return possible, instead we will follow a simple betting strategy that is to play for the outcome(win/draw/loss) that has the highest probability of occurrence according to the model's predictions and observe whether a positive return could be realized.

Chapter 2

Background

2.1 What is Football?

This part of the report will provide information regarding general principles in Football.

• Basic Rules:

– Football is one of the games where two teams play against each other. The number of players each team has is 11, and normal duration of the game is 90 minutes.

– If team A makes it on target against team B, then team A scores a goal, and earns 1 point for each goal scored.

– The team which has the highest number of points is winner. And if the points are same, then the result is assigned Draw.

• Domestic Leagues:

– On average there are 20 teams in a league, playing against each other twice in a year. Once in their own stadium, once in opponent team's stadium.

– On League level, winning a match grants a team 3 points, while losing none. And Draw results in 1 point for each team.

– The team that gains the highest number of points is declared champion In case there is more than one team reaching to the maximum point, the team with Net goals (goals scored-goals conceded) advantage is superior.

2.2 Literature Review

In this part of the project we will go over existing literature regarding model developments for prediction of match results. Applying statistical techniques on football data in order

to estimate match results has been a hot topic since previous century. Initial models proposed by Moroney (1956) [1] and Reep (1971) [2], who used Poisson and Negative Binomial distribution for modelling number of goals scored in a given match paved the way for further research around the topic. Later, coming up with teams' attacking and defensive strengths metrics, Maher [3] in 1982 used Poisson distribution to predict the expected number of goals for both home and away team to score in a match. Maher's work created the ground for Dixon and Coles [4] (1997) to further improve the Poisson model by introducing certain adjustments such as introducing correlation between home goals and away goals and assigning more weights to the draw results for low scoring games.

The Dixon-Coles model is yet to be considered as superior to the Basic Poisson model and in this project both Basic Poisson and Dixon-Coles models will be applied. Rue and Salveson [5] (2000) adjusted teams' attacking and defensive strengths by introducing time dependency, arguing that teams' performances do not stay stable over time, rather changes throughout a given season. Later, their work had been adjusted by Crowder et al. [6] (2002) so that the algorithm for extracting model coefficients work more efficiently.

In early 2000s direct approaches, rather than goals models, for prediction of soccer results attracted attention by researchers. In other words, modelling match outcomes in a direct way (win-draw-loss), instead of using goals model to estimate match scores and then obtaining respective probabilities, was preferred. For instance, Forrest and Simmons (2000) [7] used a classifier model to directly predict the match result rather than predicting the goals scored by each time. Such an approach indeed allows one to get rid of inter-dependency between the number of goals each team scores, it rather treats goals scored in home game and away game as two different variables.

Goddard [8] (2005) used an probit link regression model to predict match results by including more variables than do Basic Poisson model. In doing so, he added explanatory variables such as geographical distance between home and away teams, match significance and etc. His work was one of the pioneers to have more variables added into past data than only match results, and concluded that there was a possibility to make a positive

return when comparing his model outcomes with market odds.

Another algorithm utilized by Hamadani [9] (2006) for prediction of American football results, a different type of sport, came into play. He used Logistic Regression and SVM kernel method for results prediction.

Coming to 2010s, Adam (2016) [10] benefited from gradient descent boosting algorithm, one of the widely used optimization methods, when building a Generalized Linear Model that was applied on data that included more features than simple match results for making match outcome predictions. His work concluded importance of addition of more features, which raised model performance in his work. Certainly, what is meant by addition of more features is those features which help increase model performance.

By adding player statistics alongside historical match results for teams competing in a tournament, Tavakol (2016) [11] took Adam's approach to a different level. However, knowing that including player statistics for each team will add 22 (11 player for each team) features into Linear Model and consequently potential so-called overfitting problem could become inevitable, he decided to apply feature selection techniques to reduce dimensionality space.

When it comes to reduce feature space in a given data, there are couple of techniques previously used by statisticians. For example, Kampakis [12] made predictions for cricket matches referring to hierarchical feature design. Additionally, Tax et al. [13] (2015) made predictions on a Dutch football league by incorporating dimensionality reduction techniques with Machine Learning algorithms. Their work came to the conclusion that best results could be obtained with Principal Component Analysis (PCA) reduction algorithm incorporated with Gaussian Naïve Bayes or Perceptron classifier.

All being said, we will focus on indirect approaches- Poisson and Dixon-Coles models, from past research. And data we will refer to build our model will be historical match results in English Premier League from season 2010-2011 to season 2017-2018, and then we will observe the model performance over the data that includes matches from 2018-2019 season.

Chapter 3

Methodology

3.1 Generalized Linear Models

In this chapter we will be introducing appropriate methodology that we will be referring to when building our models for match outcome prediction. Specifically speaking, we will describe Poisson and Dixon-Coles models in this section.

3.1.1 Linear Models

A linear model describes how the response/dependent variable depends on the explanatory/independent variables. In other words, a linear model captures the linear dependency between p explanatory variables x_{1i},....,x_{pi} for *i=1,....,N* and dependent variable Y_i again for *i=1,....,N*. Furthermore, underlying assumption of the linear model is that the response variable Y_i is normally distributed. Mathematically, a linear model could be formulated as follows:

$$Y_i = \beta_0 + \beta_1 x_{1i} + ... + \beta_p x_{pi} + \epsilon_i = \mathbf{x_i^T}\beta + \epsilon_\mathbf{i}, for\ \ i=1,....,N$$

The term ϵ_i represents the difference between model predicted $\widehat{Y_i}$ and actual Y_i $for\ i = 1, ..., N$. That being said, ϵ will be referred as error term and is normally distributed with mean μ being equal to zero and variance σ^2: $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$. We could express the linear equation in matrix notation as follows:

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon,$$

where

$$\mathbf{y} = \begin{bmatrix} \mathbf{Y_1} \\ \mathbf{Y_2} \\ . \\ \mathbf{Y_N} \end{bmatrix}, \mathbf{X} = \begin{bmatrix} \mathbf{X_1^T} \\ \mathbf{X_2^T} \\ . \\ \mathbf{X_N^T} \end{bmatrix}, \beta = \begin{bmatrix} \beta_\mathbf{1} \\ \beta_\mathbf{2} \\ . \\ \beta_\mathbf{m} \end{bmatrix}, \epsilon = \begin{bmatrix} \epsilon_\mathbf{1} \\ \epsilon_\mathbf{2} \\ . \\ \epsilon_\mathbf{N} \end{bmatrix}$$

Modelling the parameter of interest-$\mathbf{Y}$ through the given independent variable-$\mathbf{X}$ with the assumption that $\mathbf{Y}$ and $\mathbf{X}$ are linearly related is also known as Linear Regression. Using historical data one could observe the dependency between $\mathbf{Y}$ and $\mathbf{X}$ and then put that into equation, thanks to which, predictions-$\widehat{Y_i}$ could be made given values of $\mathbf{X}$. And it is the term β that is representing the degree of dependency between $\mathbf{Y}$ and $\mathbf{X}$. At this stage, one should come up with β value that would minimize the error rate ϵ. That being said, there are two famously known methods that find the optimal β value for given $\mathbf{Y}$ and $\mathbf{X}$ values, which are Maximum Likelihood Estimator and Least Squares estimator. For a linear model, both of these estimators are same and can be represented as follows:

$$\widehat{\beta_{\mathbf{MLE}}} = \widehat{\beta_{\mathbf{LS}}} = (\mathbf{X^T X})^{-1}\mathbf{X^T Y}$$

Derivation of maximum likelihood and least squares estimators are outside of the scope of this project, yet one could find the derivation procedure in Dobson [14].

3.1.2 Generalized Linear Models

Although very useful and widely applied, linear model is not capable of modelling parameter of interest in every circumstances. That is due to the inherent assumption linear model makes regarding distribution of the dependent variable - $\mathbf{Y}$. In linear modelling, fundamental assumption is response variable - $\mathbf{Y}$ is normally distributed with constant mean μ and variance σ^2. However, this does not necessarily hold in every situation. For instance, there cases where parameter of interest is a binary or a count variable and there are even cases where variance of response variable depends on the mean, and thus assumption that response variable is normally distributed might not be the best one.

Having all said, it is generalized linear model, which, as its name suggests, extends the concept of a linear model to a more general form. And when applying generalized linear models the parameter of interest is allowed to be of any member from exponential family, and thus, assumption regarding normal distribution of response variable does not have to hold anymore. When talking about generalized linear model, it is important to concentrate on three main parts: Random Component, Systematic Component and Link function.

• Random Component identifies the response variable. Response variable is a random variable $\mathbf{Y}$ whose distribution depends on only parameter θ. Assuming that the probability mass function of the distribution can be expressed as follows, then the distribution belongs to exponential family:

$$f(y;\theta;\phi) = exp(\frac{y\theta - b(\theta)}{\phi} + c(y,\phi)) \tag{3.1}$$

In the exponential family, θ is the canonical parameter that depends on the model of linear predictors, which we will be talking about very soon. The term $b()$ is real-valued twice differentiable function of θ. ϕ is called the dispersion parameter and is known. Lastly, the function $c()$ is known and is independent of canonical parameter θ. And now referring to [15] we will introduce the following **Lemma**, thanks to which we will be familiarized with the mean and variance of a distribution belonging to exponential family:

Lemma 1 *If the distribution of random variable Y belongs to exponential family (3.1), $Y \sim \varepsilon$ it can be shown that:*
• Expected value of Y is equal to the first derivative of b, where b is twice differentiable function, with respect to θ :$E[Y]=b'(\theta)$.
• Variance of Y is the product of the second derivative and the scale parameter ϕ:$Var[Y]=\phi b''(\theta)$

• Systematic Component identifies the set of explanatory variables in the model. And Linear Predictor is a function of explanatory variables in linear combination with beta

values, also represented with the letter eta. In other words, the linear predictor η depicts the linear combination of X values with β values.

$$\eta = X_j^T \beta \tag{3.2}$$

• Link Function g is linking the expected value of response variable - $\mathbf{Y}$ to the linear predictor η. The link function must be monotone and differentiable:

$$g(E[Y]) = \eta$$

Definition: A Link function is called canonical if relates the canonical parameter θ directly to the linear predictor η.

$$\theta = g(b(\theta)') = \eta = X_j^T \beta$$

All being said, Generalized Linear Models can be used in modelling count/binary variables as well, if and only if a distribution is a member of exponential family of distributions. Using Generalized Linear Models we refer to Maximum Likelihood method for estimating model parameters - $\beta_1, ...\beta_p$. In the following section we will introduce models that we will be referring to during our analysis.

3.2 Bernoulli Trials and Binomial Distribution

In this section, we will talk about Bernoulli trials, Binomial Distribution and how they pave the way for understanding of one of the most famous distributions, namely Poisson Distribution. Let us consider an experiment where we will toss a coin and the result is either "Heads" or "Tails", denoted by "H" and "T" respectively. Assuming that probability of "Heads" to show up is p and $p \in (0, 1)$, one can come up with probability of "Tails" to occur as q where $q = 1 - p$, since these two events are assumed to be mutually exclusive. This is called a Bernoulli Trial with probability of "Heads" to occur being equal to p and "Tails" being equal to q. Now we define a Bernoulli random variable $\mathcal{S}$, where $\mathcal{S} = \{\mathcal{H}, \mathcal{T}\}$ and $X : H \to \{0, 1\}$ be a function on the sample space $\mathcal{S} : X(H) = 1$ and $X(T) = 0$. Then:

$$P_X(0) = P(X = 0) = q = 1 - p$$

$$P_X(1) = P(X = 1) = p$$

Thus the random variable X is referred as Bernoulli Random Variable with probability of occurrence of "H" being equal to p, and probability mass function of $\mathbf{P_X}$. Now that we have defined Bernoulli random variable, let's now assume that one performs $\mathbf{n}$ independent Bernoulli Trials with number of successes being represented by Y, and success occurs each time the coin turns up as "Heads". Knowing that probability of heads turning up is p, then for $0 \leq x \leq n$, where $x \in \mathbf{Z}$, we have:

$$P_y(x) = P(Y = x) = \binom{n}{x} p^x (1-p)^{n-x} = \binom{n}{x} p^x q^{n-x} \tag{3.3}$$

One can attempt to understand the formula above in an intuitive sense. Since p represents the probability of a coin turning up "Heads", and assuming that all trials are made independently, p^x shows the probability of "Heads" landing x times and since the "Heads" turning up x times out of n trials imply that "Tails" should turn up $n - x$ times, its probability of occurrence translates into $(1-p)^{n-x}$ or $(q)^{n-x}$. The last thing to understand in our formula is $\binom{n}{x}$, which shows number of possible sequences where "Heads" will occur exactly x times out of n trials. And finally, the random variable Y is called Binomial Random Variable with probability mass function being equal to $P_y(x)$, with parameters n and p. At this stage, one could go further and derive expectation of random variable Y being equal to np and variance being equal to npq, yet the derivation procedure is outside of the scope of this project work.

3.3 Poisson Distribution

Now that we have introduced Binomial Random Variable, it would be important to note that for computing probabilities for a binomial random variable Y, it is comfortable to work with relatively small number of trials, or for a small n. For example, imagine that

one wants to know the number of cars pass from a given street in a day, and then one specifies the time interval to be an hour instead of a day. And then one could obviously try to find the probability of certain number of cars passing from a street in a second, and this translates into increase in the number of trials being made. Assuming that $X \sim \mathcal{B}(n,\, \lambda/n)$ where $\lambda > 0$, and relying on our definition of p and q in **section 3.2** then:

$$
\begin{aligned}
P_n(x) &:= \frac{n!}{x!(n-x)!}(p_n)^x(q_n)^{(n-x)} \\
&= \frac{n!}{k!(n-x)!}(\frac{\lambda}{n})^x(1-\frac{\lambda}{n})^{n-x} \\
= \frac{n(n-1)(n-2)....(n-x+2)(n-x+1)}{x!}&(\frac{\lambda}{n})^x(1-\frac{\lambda}{n})^{n-x} \\
= \frac{n(n-1)(n-2)....(n-x+2)(n-x+1)}{n^x}&(\frac{\lambda^x}{x!})\frac{(1-\frac{\lambda}{n})^n}{(1-\frac{\lambda}{n})^x}
\end{aligned}
$$

Now as we said in the beginning, taking the number of trials go infinity, or in other words, taking the limit as $n \to \infty$, we have:

$$
lim_{n\to\infty}P_n(x) \to 1\frac{\lambda^x}{x!}\frac{e^{-\lambda}}{1} = e^{-\lambda}\frac{\lambda^x}{x!} \stackrel{\text{def}}{=:} f(x;\lambda) \tag{3.4}
$$

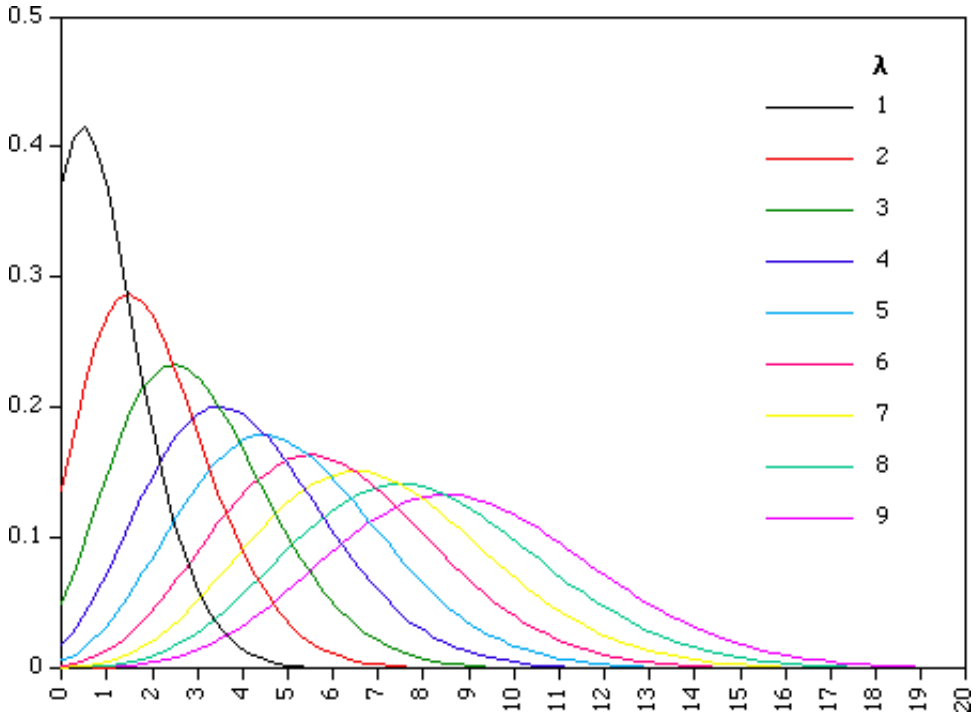
With that being said, if n is large enough and correspondingly, λ/n is small enough, then (3.4) can be used to approximate (3.3). Moreover, a random variable whose probability mass function can be expressed as in (3.4) is called Poisson random variable with parameter λ. Now that we have talked about the logic and derivation of Poisson Random Variable, we will be talking about Poisson distribution in the context of Generalized Linear Models.

Named after the famous french mathematician *Siméon Denis Poisson*, Poisson Distribution is a non-negative discrete probability distribution. Poisson distribution was first published in 1837 in his work - *'Recherches sur la probabilité des jugements en matières criminelles et matière civile'* [16].

The Poisson Distribution measures the probability of a certain number of events occurring in a given fixed amount of time, with a condition of such events occurring

independently of one another with a known constant average rate that we refer as λ. The parameter λ is the rate parameter of the distribution and is equal to the average number of events happening in a fixed interval of time.

Figure 3.1: Poisson Distribution



Coming to the formal definition of Poisson Distribution, the random variable Y is said to follow the Poisson Distribution with a known rate parameter $\lambda > 0$, if for $y = 0, 1, 2...$, the probability mass function of Y can be expressed in the following way:

$$f(y;\lambda) = P(Y = y) = \frac{\lambda^y e^{-\lambda}}{y!} \tag{3.5}$$

And referring to (3.1) one could show that Poisson Distribution belongs to exponential family of distributions in the context of Generalized Linear Models.

$$f(y;\theta;\phi) = \frac{e^{-}\lambda\lambda^y}{y!} = exp(log(e^{-\lambda}) + log(\lambda)^y - log(y!))$$

$$\Rightarrow f(y;\theta;\phi) = exp(ylog(\lambda) - \lambda - log(y!)) \tag{3.6}$$

$$\Rightarrow \phi = 1; \quad \theta = log(\lambda) \Leftrightarrow \lambda = e^{\theta} \quad b(\theta) = \lambda = e^{\theta} \quad c(y;\phi) = -log(y!)$$

Referring to **Lemma 1**, one could express the expectation and variance of Poisson distribution as follows:

$$E[Y] = b(\theta)' = e^{\theta} = \lambda \quad Var[Y] = b(\theta)''\phi = e^{\theta} = \lambda$$

With that being said, it is now easily seen that expectation of Poisson Distribution is equal to its variance, which is one of the characteristics for Poisson Distribution. And since $\theta = log(\lambda)$, or in other words, log-link of the expectation of Poisson random variable is canonical, we say that one of the link functions that maps expectation to linear predictor is log-link, and thus, we will choose log-link to continue:

$$\theta = log(b(\theta)') = \eta = log(\lambda) = X_j^T \beta \tag{3.7}$$

3.4 Poisson Regression

Poisson regression model is a generalized linear model used to model count data. As said earlier, the parameter of interest to be modeled could be continuous, binary or count and for the purpose of this project work, we are interested in modelling count data, which are number of goals in our analysis. The Poisson regression model is derived from Poisson distribution with rate parameter λ depending on the explanatory variables, which are teams in our analysis.

Data that is used in Poisson regression model consists of sample of N observations with independent response variable - Y_i and explanatory variables - x_i for $i = 0, 1, ..., N$. The response variable Y_i is the number of occurrences of a given event, whereas x_i is the vector of linearly independent explanatory variables that are supposed to be determining the response variable. And we will build a regression model by conditioning Y_i on a p-dimensional vector $x_i^T = [x_{1i}, x_{2i}....x_{pi}]$ and coefficient parameters $\beta = [\beta_1, \beta_2, ...\beta_p]$ such that $E[y_i|x_i] = \lambda_i(x_i, \beta)$. In the **section 3.4**, more specifically in (3.7) we already proposed that we will be choosing logarithmic link function that is mapping the expectation of Poisson random variable into linear predictor- η_i. Then, more formally, one could define the Poisson regression equation in the following

way:

$$f(y_i|x_i) = \frac{\lambda_i^{y_i} e^{-\lambda_i}}{y_i!} \quad i = 1, 2, ..., N \tag{3.8}$$

with

$$\lambda_i = e^{x_i^T \beta}$$

Now that we are left with derivation of β estimates, we will spare the next section of our project work on the derivation procedure of our β parameters. At this stage, one should note that since we know probability mass function of Poisson Distribution it would be reasonable to rely on Maximum Likelihood Method for finding β estimates. Additionally, one has to use iterative re-weighted least squares algorithm to solve the system of equations, also called as score equations that we will be talking about in the next section.

3.5 Parameter Estimation Procedure

Referring to statistical literature, especially to [14] we will be building our procedure of estimating β parameters for Poisson regression.

Now let's consider independent random variables Y_1, Y_2, ..., Y_N, which are Poisson distributed with rate parameter λ_i for $i = 1, 2, ..., N$. And referring to (3.8), we already got ourselves familiarized with the notion that Poisson distribution satisfies the properties of generalized linear models. Then, to estimate parameter vector β one needs to apply maximum likelihood method and arrive at parameter coefficients that maximize the likelihood of generating the observations given the parameters. And in our example, β values and response variables $Y_i's$ are related through λ_i values, and we already know that $E[\,Y_i] = \lambda_i$. More specifically, it is logarithmic link function that maps λ_i values to linear predictor or since $\lambda_i = e^{x_i^T \beta}$, we can write $g(E[\,Y_i]) = log(\lambda_i) = x_i^T \beta$, where x_i is an p-dimensional vector of explanatory variables $x_i^T = [x_{1i}, x_{2i}....x_{pi}]$. Our Likelihood function for each Y_i can be given as follows:

$$L(\theta_i; y_i) = f(y_i; \theta_i; \phi) = exp(\frac{y_i\theta_i - b(\theta_i)}{\phi} + c(y_i, \phi)) \tag{3.9}$$

And correspondingly, our log-likelihood function is logarithmic transformation of (3.9):

$$l_i(\theta_i) = lnL(\theta_i; y_i) = \frac{y_i\theta_i - b(\theta_i)}{\phi} + c(y_i, \phi) \tag{3.10}$$

where $b()$, $c()$, $a()$ are functions that we will be replacing with exponential family form of Poisson distribution. Referring back to (3.6), one could plug in the values in the equation into those in (3.10) and get the following:

$$l_i(\lambda_i; y_i) = y_i log(\lambda_i) - \lambda_i - log(y_i!) \tag{3.11}$$

And since (3.11) represents the log-likelihood of only one response variable and variables Y_i are independent, the log-likelihood of the whole sample over $y_1, ...y_N$ could be represented as follows:

$$l(\lambda; \mathbf{y}) = \sum_{i=1}^{N} l_i(\lambda_i; y_i) = \sum_{i=1}^{N} y_i log(\lambda_i) - \sum_{i=1}^{N} \lambda_i - \sum_{i=1}^{N} log(y_i!)$$

At this stage, one needs to come up with some way that the equation above includes β terms. In order to arrive at maximum likelihood estimate for β parameters, we need score function.

$$S_j = \frac{\partial l}{\partial \beta_j} = \sum_{i=1}^{N} \frac{\partial l_i}{\partial \beta_j}$$

and using chain rule:

$$= \sum_{i=1}^{N} \frac{\partial l_i}{\partial \lambda_i} \frac{\partial \lambda_i}{\partial \beta_j} \tag{3.12}$$

Using 3.11 and (3.2) and applying chain rule we have:

$$\frac{\partial l_i}{\partial \lambda_i} = \frac{y_i}{\lambda_i} - 1$$

and

$$\frac{\partial \lambda_i}{\partial \beta_j} = \frac{\partial \lambda_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_j} = \frac{\partial \lambda_i}{\partial \eta_i} x_{ij}$$

and since $\frac{\partial \lambda_i}{\partial \eta_i} = \lambda_i$ referring to (3.8), we could write previous equation as follows:

$$\frac{\partial \lambda_i}{\partial \beta_j} = \frac{\partial \lambda_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \beta_j} = \frac{\partial \lambda_i}{\partial \eta_i} x_{ij} = \lambda_i x_{ij} = e^{x_i^T \beta} x_{ij}$$

Coming back to our initial point of estimating β values, one should note that there is no direct estimation method, which is why we will refer to (3.13) that starts from m^{th} iteration for finding vector of β values:$(\beta_1, \beta_2, ..., \beta_p)$. Now at this stage, we will skip the derivation procedure of how one can reach at (3.13) given our explanation above regarding parameter estimation. (3.13) is obtained with the help of Newton-Raphson formula alongside with method of scoring, where initial guess for the parameter of interest is made and then successive approximations are obtained. One can refer to [14] for more detailed explanation and be familiarized with the procedure Newton-Raphson formula is implemented :

$$b^{(m)} = b^{(m-1)} + [\mathcal{J}^{(m-1)}]^{-1} S^{(m-1)} \tag{3.13}$$

And one should certainly note that the term b^m represents the m^{th} iteration estimate for β vector. In (3.13), $[J^{(m-1)}]^{-1}$ is the inverse of **information matrix** $\mathcal{J}_{jk}$ and $S^{(m-1)}$ is the vector of elements, all being evaluated at $\mathbf{b^{(m-1)}}$. Multiplying both sides of (3.13) by $\mathcal{J}^{(m-1)}$, one obtains:

$$\mathcal{J}^{(m-1)} b^{(m)} = \mathcal{J}^{(m-1)} b^{(m-1)} + S^{(m-1)} \tag{3.14}$$

And now, we could expand the terms in equation (3.14) referring to [16]. In short, right hand-side of (3.14) could be written as:

$$\sum_{k=1}^{p}\sum_{i=1}^{N} \frac{x_{ij}x_{ik}}{\lambda_i}(\frac{\partial \lambda_i}{\partial \eta_i})^2 b_k^{(m-1)} + \sum_{i=1}^{N}(\frac{y_i - \lambda_i^{m-1}}{\lambda_i^{m-1}})(\frac{\partial \lambda_i^{m-1}}{\partial \eta_i})x_{ij}$$

again since $\frac{\partial \lambda_i}{\partial \eta_i} = \lambda_i$:

$$\sum_{k=1}^{p}\sum_{i=1}^{N} x_{ij}x_{ik}\lambda_i b_k^{(m-1)} + \sum_{i=1}^{N}(\frac{y_i - \lambda_i^{m-1}}{\lambda_i^{m-1}})(\frac{\partial \lambda_i^{m-1}}{\partial \eta_i})x_{ij}$$

being evaluated at $\mathbf{b^{(m-1)}}$. Therefore, we could also write:

$$\mathbf{X^T W z}, \tag{3.15}$$

where

$$z_i = \sum_{k=1}^{p} x_{ik} b_k^{(m-1)} + (y_i - \lambda_i)(\frac{\partial \eta_i}{\partial \lambda_i})$$

again since $\frac{\partial \lambda_i}{\partial \eta_i} = \lambda_i$:

$$z_i = \sum_{k=1}^{p} x_{ik} b_k^{(m-1)} + \frac{(y_i - \lambda_i)}{\lambda_i}$$

with λ_i being evaluated at $\mathbf{b^{(m-1)}}$. And finally, the iterative method (3.14) could be rewritten as follows:

$$\mathbf{X^T W^{(m-1)} X b^{(m)}} = \mathbf{X^T W^{(m-1)} z^{(m-1)}}. \tag{3.16}$$

where (3.16) is in its linear representation form, where $\mathbf{W}$ and $\mathbf{z}$ depend on $\mathbf{b}$. And this way, we have approximated parameter estimates-β vector using iterative re-weighted least squares algorithm. However, it should be noted that in the software we will be referring to(R and Python 3.0), there will not be a need for following the procedure for parameter estimation since there are already built-in functions doing the parameter estimation task for us.

3.6 Dixon-Coles Model

The other model we will be applying during our analysis is famously known as Dixon-Coles model. Initially developed by Mark J.Dixon and Stuart G.Coles, Dixon-Coles model is considered as alternative for Basic Poisson model. Yet, before starting to describe the model itself, we will justify the need for Dixon-Coles model in our project work.

3.6.1 Poisson Model Output

In the Basic Poisson model, once we have the coefficient parameters- β vector, and regression outputs extracted, the next step would be to use those regression outputs to predict

probabilities for different possible match scores. Assume we want to predict possible match outcome probabilities for any two teams playing against each other. Although we will give more information regarding model outputs in **Chapter 6**, it would be reasonable to give brief information regarding how the model works to justify the need for our usage of Dixon-Coles model.

When building Poisson model, we will have data consisting of **Team**, **Opponent** and **Home advantage** as our explanatory variables, where **Team** column specifies the teams playing, while **Opponent** column demonstrates the teams playing against those teams in **Team** column. And finally, **Home advantage** is the dummy variable: 1 if the team is playing at its own stadium, and 0 otherwise. For example, assume we have data of only 1 match between two teams playing: Arsenal and Bournemouth. In order to run Poisson regression for goals over Home Team, Opponent and Home advantage variable, we will structure data in the following way for each match:

Table 3.1: Structure of data

Team	Opponent	Home	Goals
Arsenal	Bournemouth	1	3
Bournemouth	Arsenal	0	0

What's important to realize regarding the table above, there is only one match, played between Arsenal and Bournemouth, which ended 3-0 with Arsenal declaring victory. Yet, as can be seen from the table, we double the number of rows for each match and the corresponding columns represent teams rather than matches. In this case, we will first treat Arsenal as team, and Bournemouth as opponent to the team. And since Arsenal is home team in this example, home dummy variable indicates 1 for Arsenal and response variable in this case- $\mathbf{Y_1}$ and its value is 3. The same way, we will treat Bournemouth as team and Arsenal as opponent to Bournemouth. However, since Bournemouth plays away, the column Home(indicating whether the team plays home or away) is 0 and value of response variable $\mathbf{Y_2}$ is 0 in this case. As we talked in **section 3.4** $\mathbf{Y}$ represents the column 'goals'. Moreover, $\mathbf{Y}$ consists of both home team goals(goals Arsenal scored)

and away team goals(goals Bournemouth scored).

Once we regress $\mathbf{Y}$ over Home Team, Opponent and Home advantage variables, from the model output we will get two coefficients for each team in our sample and then by exponentiating those coefficients, we will obtain values representing the attacking and defensive strength of each team. Additionally, one other value we obtain will be representing **Home advantage**, in other words it will capture the home effect, meaning that it will demonstrate the expected number of times teams playing at their own stadium will score more goals than teams playing away.In **Chapter 6** we will talk more about attacking and defensive strength of each team and how coefficients obtained from Poisson model relate to team strength measures.

Referring back to (3.5), λ in our model represents a certain team's average or expected number of goals to be scored in a given match. And since in a match there should be exactly two teams playing, we will have two different λ values, λ_{home} representing home team's expected number of goals against away team and λ_{away}, representing away team's expected number of goals against home team. However, for the sake of simplicity, we will call λ_{away} as μ and λ_{home} as λ from now on. Once we know expected number of goals for a given team to score against another team, in prediction stage over Poisson model we will plug those λ and μ values into (3.17) to extract the probabilities for potential match outcomes. The basic assumption regarding (3.17) is that the number of goals to be scored by home and away teams in any given match follow Poisson distribution and are independent. More formally, in a match between teams i and j, let $U_{i,j}$ and $V_{i,j}$ be the number of goals be scored by home and away team respectively. Then, the assumption that :

$$U_{i,j} \sim Poisson(\alpha_i \beta_j \gamma)$$

$$V_{i,j} \sim Poisson(\alpha_j \beta_i)$$

$for i, j = 1, 2, ...m, i \neq j$ and m is the number of unique teams participating in Premier League in season 2018-2019(we have explained the difference between values n and m in **section 4.1**). Again, we need the assumption of independence between the

number of goals scored by home and away teams in prediction stage, after modelling our response variable-goals.

where $U_{i,j}$ and $V_{i,j}$ are independent, and forms the main rationale to arrive at (3.17):

$$P(U_{i,j} = u;\, V_{i,j} = v) = \frac{\lambda^u e^{-\lambda}}{u!}\frac{\mu^v e^{-\mu}}{v!}, \tag{3.17}$$

where $\lambda = \alpha_i\beta_j\gamma$ and $\mu = \alpha_j\beta_i$, representing home and away teams' expected number of goals to be scored against each other, where $\alpha_i, \beta_j, \gamma > 0$. The subscripts i and j stand for home and away teams, and α and β values representing attack and defensive strength respectively, while γ parameter denotes home advantage coefficient. Last yet not the least,u and v values represent the number of goals each team can score and by plugging in different values for u and v, in fact, one obtains the probability of potential match score between two teams. In general, it depends on the practitioner to specify intervals for u and v values and we will rationalize our method of interval selection for different possible match scores in the coming sections.

By introducing (3.17) we have arrived at a very important point. Previously in (3.5) we defined probability mass function for Poisson distribution where we assumed only one random variable Y, which is goals in our data to be modeled by Poisson regression. In the case of (3.17), however, we have two random variables - U and V, and correspondingly, two independent Poisson distributions. That is, after modelling the variable Y, goals in our case, we will predict the probabilities of possible match scores by inserting expected goals we obtain from model output for teams that play against each other into (3.17). We will continue talking more about the model outputs in **Chapter 6**, now we could start explaining Dixon-Coles model in a more detailed sense.

3.6.2 Dixon-Coles Model

Now that we have briefly talked about the procedure of deriving probabilities from Poisson regression outputs, we could start researching potential flaws Poisson model might inherit when relying on (3.17) to extract probabilities of possible match outcomes. In fact, it is equation (3.17) where Dixon-Coles model comes into play, suggesting that independence

assumption between the number of goals scored by home and away teams might not be the best one and at certain times there could be a superior model with more realistic approach. With that being said, in their original paper, Mark J.Dixon and Stuart G.Coles argue that for the matches ending with the scores $0-0; 0-1; 1-0; 1-1$, the assumption regarding the independence of number of goals scored by home and away teams is flawed. One could check [4] to follow the procedure on proof of existing correlation between number of goals scored by home and away teams when match ends with low scores. Eventually, in order to provide a potential alternative for (3.17), they proposed the following approach:

$$P(U_{i,j} = u; V_{i,j} = v) = \tau_{\lambda,\mu}(u, v)\frac{\lambda^u e^{-\lambda}}{u!}\frac{\mu^v e^{-\mu}}{v!}, \tag{3.18}$$

where

$$\lambda = \alpha_i \beta_j \gamma$$

$$\mu = \alpha_j \beta_i$$

and

$$\tau_{\lambda,\mu}(u, v) = \begin{cases} 1\text{-}\lambda\mu\rho & \text{if u=v=0,} \\ 1\text{+}\lambda\rho & \text{if u=0, v=1,} \\ 1\text{+}\mu\rho & \text{if u=1, v=0,} \\ 1\text{-}\rho & \text{if u=1, v=1,} \\ 1 & \text{otherwise.} \end{cases}$$

In the function above, ρ parameter is also referred as correction, controlling the dependence between goals scored by home and away teams when the match ends with the results $0-0; 0-1; 1-0; 1-1$. One could easily check that once the ρ parameter is equal to 0, then Dixon-Coles model and Poisson model are equal. In our case, Dixon-Coles model specified the interval for ρ parameter to be as following:

$$max(-1/\lambda, -1\mu) \leq \rho \leq min(1/\lambda\mu, 1)$$

Now before we move any further, we will show that (3.18) is a valid probability distribution, in other words, we will have to prove that:

$$\sum_{u=0}^{\infty} \frac{\lambda^u e^{-\lambda}}{u!} \sum_{v=0}^{\infty} \frac{\mu^v e^{-\mu}}{v!} \tau_{\lambda,\mu}(u,v) = 1$$

And we will first prove that

$$\sum_{u=0}^{\infty} \frac{\lambda^u e^{-\lambda}}{u!} \sum_{v=0}^{\infty} \frac{\mu^v e^{-\mu}}{v!} \tag{3.19}$$

is a valid probability distribution. Expanding (3.19), we have:

$$e^{-\lambda}e^{-\mu}(1 + \frac{\lambda}{1!} + \frac{\lambda^2}{2!} + ...(1 + \frac{\mu}{1!} + \frac{\mu^2}{2!} + ...)) = e^{-\lambda}e^{-\mu}e^{\lambda}e^{\mu} = 1$$

Since we proved (3.19) is a valid probability distribution, we could expand it by moving the summation signs, and we already know that (3.19) adds up to 1:

$$\sum_{u=0}^{\infty}\sum_{v=0}^{\infty} \frac{\lambda^u e^{-\lambda}}{u!}\frac{\mu^v e^{-\mu}}{v!} = e^{-\lambda}e^{-\mu}(1 + \mu + \lambda + \lambda\mu + \\ + \frac{\lambda^2}{2!} + \frac{\mu^2}{2!} + \frac{\lambda\mu^2}{2!} + \frac{\mu\lambda^2}{2!} + ..) = 1$$

And now, we can prove that (3.18) is a valid probability distribution as well:

$$\sum_{u=0}^{\infty} \frac{\lambda^u e^{-\lambda}}{u!} \sum_{v=0}^{\infty} \frac{\mu^v e^{-\mu}}{v!} \tau_{\lambda,\mu}(u,v) =$$

$$= e^{-\lambda}e^{-\mu}(1 - \lambda\mu\rho + \mu + \lambda\mu\rho + \lambda + \lambda\mu\rho + \lambda\mu - \lambda\mu\rho + \\ + \frac{\lambda^2}{2!} + \frac{\mu^2}{2!} + \frac{\lambda\mu^2}{2!} + \frac{\mu\lambda^2}{2!} + ...)$$

And simplifying the equation, we obtain:

$$= e^{-\lambda}e^{-\mu}(1 + \mu + \lambda + \lambda\mu + \frac{\lambda^2}{2!} + \frac{\mu^2}{2!} + \frac{\lambda\mu^2}{2!} + \frac{\mu\lambda^2}{2!} + ...) = 1$$

which is the same as the (3.19) and thus must add up to 1.

Regarding parameter estimation, under Dixon-Coles model, there are $2n + 2$ number of parameters to be estimated, namely, n number of α values: $[\alpha_1, ...\alpha_n]$, representing attacking strength, and n number of β values: $[\beta_1, ...\beta_n]$, denoting defensive strengths

and additionally, ρ parameter introducing dependency and γ parameter denoting home advantage. Moreover, in our case, there is one constraint being added into the computation of parameters so that the model we build will not suffer from overparametrization:

$$\frac{1}{n}\sum_{i=1}^{n}\alpha_i = 1,$$

where $\alpha_i \geq 0$

So that average attacking strength would be equal to 1. This step makes sure that the model does not suffer from overparametrization, yet one could refer to [4] for additional information regarding constraints added into our model.

Given the constraint, we want to find those $2n + 2$ parameters that would maximize the following likelihood function:

$$L(\alpha_i, \beta_i, \rho, \gamma; i = 1, ...n) = \prod_{i,j;i\neq j}^{n} \tau_{\lambda_{i,j},\mu_{i,j}}(u_{i,j}, v_{i,j}) \frac{\lambda_{i,j}^{u_{i,j}} e^{-\lambda_{i,j}}}{u_{i,j}!} \frac{\mu_{i,j}^{v_{i,j}} e^{-\mu_{i,j}}}{v_{i,j}!},$$

However, one should consider the fact that the equation above assumes only two matches between each team. In our case we have sampled matches from season 2010-2011 up to 2017-2018, and thus the number of matches between teams is more than two. In order to take this fact into consideration, we will introduce certain adjustment to the equation above. Let $k(i, j)$ be the number of matches played between team i and team j in case i is home team, then:

$$\sum_{i,j;i\neq j}^{n} k(i,j) = k,$$

where k is the total number of matches being played during our sample of train data from season 2010-2011 to 2017-2018. And since we double number of matches in our data as we discussed in **subsection 3.6.1**, the total number of observations N=2k. Then

we could write the adjusted likelihood function as following:

$$L(\alpha_i, \beta_i, \rho, \gamma; i = 1, ...n) = \prod_{i,j;i\neq j}^{n} \prod_{l=1}^{k(i,j)} \tau_{\lambda_{i,j},\mu_{i,j}}(u_{i,j}^l, v_{i,j}^l) \frac{\lambda_{i,j}^{u_{i,j}^l} e^{-\lambda_{i,j}}}{u_{i,j}^l!} \frac{\mu_{i,j}^{v_{i,j}^l} e^{-\mu_{i,j}}}{v_{i,j}^l!}, \tag{3.20}$$

where

$$\begin{aligned} \lambda_{i,j} &= \alpha_i \beta_j \gamma, \\ \mu_{i,j} &= \alpha_j \beta_i \end{aligned} \tag{3.21}$$

where $i, j = 1, ...(n)$ and $i \neq j$, are the indexes for home and away teams and again $k(i, j)$ is the number of matches played between team i and team j, in cases i is home team, and $(u_{i,j}^l, v_{i,j}^l)$ denote the home and away goals for l-indexed match. Additionally, i and j be the indices of corresponding home and away teams respectively, then it is (3.20) that we want to maximize given (3.21). Taking log of likelihood function (3.20), and assuming that each match is independent from one another, we have:

$$\begin{aligned} l(\alpha_i, \beta_i, \rho, \gamma; i = 1, ...n) = \sum_{i,j;i\neq j}^{n} \sum_{l=1}^{k(i,j)} log(\tau_{\lambda_{i,j},\mu_{i,j}}(u_{i,j}^l, v_{i,j}^l)) + \\ u_{i,j}^l log(\lambda_{i,j}) - \lambda_{i,j} - \lambda_{i,j} u_{i,j}^l! + v_{i,j}^l log(\mu_{i,j}) - \mu_{i,j} - \mu_{i,j} v_{i,j}^l! \end{aligned} \tag{3.22}$$

In our python code, we maximize the equation above by adding teams' attacking strength constraint. Additionally, our procedure is as follows:

-Specify initial attacking strengths for each team randomly between 0 and 1.

-Specify initial defensive strengths for each team randomly between 0 and 1.

-Specify initial dependence-ρ parameter as 0 (Or assume no dependence initially).

-Specify initial home advantage-γ parameter as 1.

-Add following constraint:

$$\frac{1}{n} \sum_{i=1}^{n} \alpha_i = 1$$

-Numerically maximize (3.22).

And in order to make this process easier, we will use minimize function from scipy.optimize package, and then minimize the negative log-likelihood or negated version of (3.22). After first iteration of the procedure detailed above through randomly assigning initial parameters for those 2n+2 parameters we obtain first 2n+2 coefficients for those parameters at first iteration. And those coefficients relate to our α, β, γ values as follows:

$$\tilde{\alpha_i} = log\alpha_i \quad \tilde{\alpha_j} = log\alpha_i \quad \tilde{\beta_i} = log\beta_i \quad \tilde{\beta_j} = log\beta_j \quad \tilde{\gamma} = log\gamma$$

Then, we could write (3.21) in the following way:

$$\lambda_{ij} = \alpha_i \beta_j \gamma = e^{\tilde{\alpha_i} + \tilde{\beta_j} + \tilde{\gamma}}$$

$$\mu_{ji} = \alpha_j \beta_i = e^{\tilde{\alpha_j} + \tilde{\beta_i}}$$

where home advantage $\tilde{\gamma}$ is added only in calculation of λ values.

Now that we have described the models we will be applying in our analysis, we could safely switch to talk about the data we will be working with.

Chapter 4

Data

4.1 Premier League

In our research we will be focusing on teams participating in Premier League in England. Premier League is the highest division football league in England and there are 20 teams competing each season. We will be working for giving predictions for matches played in the season 2018-2019 using historical data from the season 2010-2011 up until the season 2017-2018 included.

Talking about the teams, as said there are 20 teams competing in the league and each team plays 38 games in a given season: 19 away and 19 home, making in total 380 matches among teams. We will be taking the sample of 8x380=3040 matches to build our model and then give predictions for the 380 matches played in season 2018-2019. In other words, we will be taking training data as matches played from season 2010-2011 till season 2017-2018 and test data as matches played during season 2018-2019 in Premier League.

One should also note that at the end of each season 3 worst performing teams relegate to the second division, while 3 best performing teams from second division gain right to perform in Premier League. With that being said, we sampled our data so that those teams that compete in the season 2018-2019 are already included in our sample data, so that our model would make predictions for every game played in 2018-2019.

Table 4.1: Number of Teams

n	Number of unique teams from season 2010-2011 to season 2017-2018
m	Number of unique teams in season 2018-2019

Another important thing to take from our data is that in this project n represents number of unique teams performed in Premier League from season 2010-2011 up to season 2017-2018. On the other hand, m represents number of unique teams performed in Premier League in season 2018-2019. Since three team relegates to the second division and three new teams replace those that got relegated each season $n > m$. Specifically speaking, in our data n is 35 while m is 20.

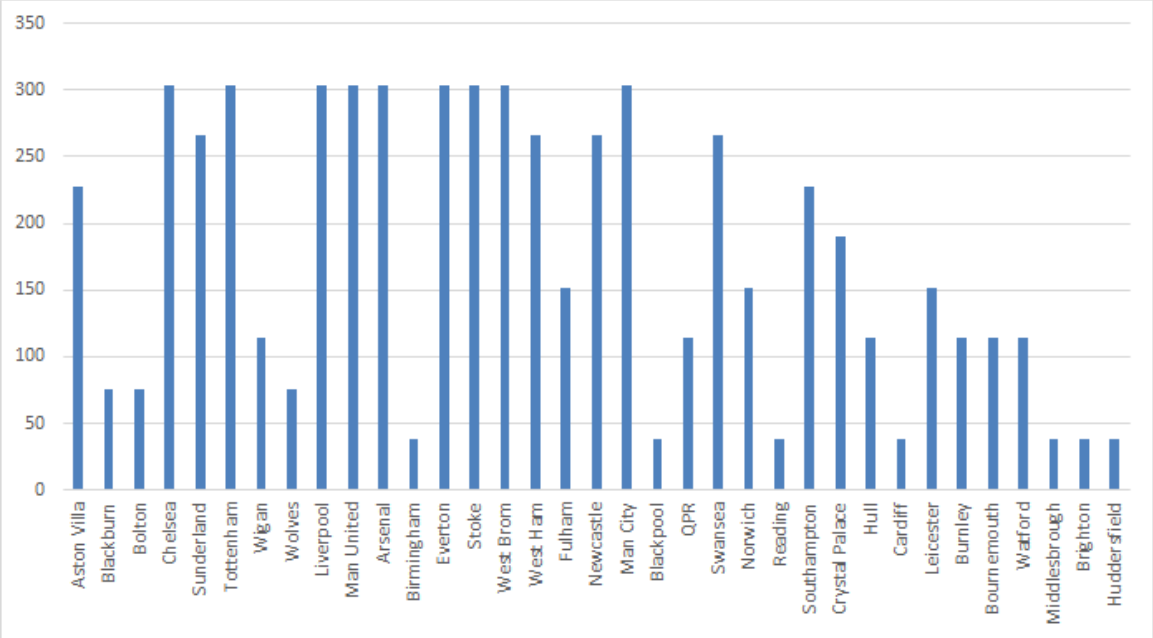
4.2 Preliminary Statistics over Data

Talking about data, it is important to emphasize that one could easily access data regarding match results from various reliable sources over internet. The source we referred to is *football-data.co.uk*[17]. The most important information for us from those tables in *football-data.co.uk* would be categorized as following:

- Home Team
- Away Team
- Home Team goals
- Away Team Goals
- Bwin win odds for home team
- Bwin draw odds
- Bwin win odds for away team

Once we have this information at hand, then we will be ready to start our analysis. However, one should certainly note that because of relegation and addition of three teams into first division each season, number of games each teams play are different. Although we have all the teams in our sample we want to predict for 2018-2019, there is still discrepancy in the number of games each team plays. One could observe this in the Figure 4.1 below:

Figure 4.1: Number of matches each team played from 2010-2011 up to 2017-2018 in Premier League



Chapter 5

Betting Strategy

In this chapter, we will give information regarding the potential betting strategies one might refer to when trying to receive positive return over investment in a certain match. More specifically, potential betting strategies will be proposed and at the end of our analysis, such strategies will be compared so that one could realize the most effective investment strategy for investment in Premier League match outcomes in the season 2018-2019. However, before talking about the possible investment strategies, we will introduce the concept of risk-free investment with a positive return on betting market, also known called arbitrage strategy over betting market.

5.1 Arbitrage Strategy

In this section we will introduce a strategy thanks to which one could make profit without risking any loss at all. We will refer to such a strategy as arbitrage strategy in betting market, where no preliminary knowledge regarding team performance or expertise in the betting market is necessary. Our very purpose of talking about such strategies is to show that in the existence of arbitrage strategies one should definitely not proceed with other strategies we will be introducing in the next section, unless he/she wants is still willing to risk for some higher return. Let A, B, C be mutually exclusive events such that $(A) \cap (B) \cap (C) = \emptyset$ implying that $P(A \cap B \cap C) = 0$, and j be the index for a match being played and O be the odd provided by the bookmaker(1 divided odd for a specific event yields the probability of that specific event calculated by bookmaker), then

for every match j, if:

$$\frac{1}{O_{Aj}} + \frac{1}{O_{Bj}} + \frac{1}{O_{Cj}} < 1 \tag{5.1}$$

then we say one could make riskless return through betting on match j. It would be fairly easy to prove (5.1). For the sake of simplicity, assume two mutually exclusive events in a match: event A and event B. In the case of football betting event A might be thought of the match ending above 2.5 goals and event B otherwise. Again assume that we have odds indexed by A and B: O_A representing the odd for event A to happen, while odd O_B representing the odd for event B to happen. One chooses to bet a amount on event A and b amount on event B simultaneously. Then the profit one could make can be described as follows:

$$\begin{cases} aO_A - (a+b) & \text{if event A happens,} \\ bO_B - (a+b) & \text{if event B happens.} \end{cases}$$

and since we know that

$$\begin{cases} \frac{1}{O_A} = \mathrm{P(A)}, \\ \frac{1}{O_B} = \mathrm{P(B)}, \end{cases}$$

then

$$\begin{cases} \frac{a}{P(A)} - (a+b) & \text{if event A happens,} \\ \frac{b}{P(B)} - (a+b) & \text{if event B happens.} \end{cases}$$

one may select a = P(A) and b = P(B). Then we have:

$$\begin{cases} 1 - (P(A) + P(B)) & \text{if event A happens,} \\ 1 - (P(A) + P(B)) & \text{if event B happens.} \end{cases}$$

For any case, we can see that our profit is equal to 1-(P(A)+P(B)). And as said earlier, since odds are inverse of probabilities for the events to happen, we could express our profit as $1 - (\frac{1}{O_A} + \frac{1}{O_B})$. Consequently, our profit becomes positive if and only if $(\frac{1}{O_A} + \frac{1}{O_B}) < 1$, and this proves (5.1) in the case of two events. Indeed, increasing the number of events under the assumption that those events are mutually exclusive yields in the same result. However, such opportunities in the betting market are not always

possible since (5.1) does not hold in most of the cases. In case (5.1) holds for any number events assuming that those events are mutually exclusive, then one should definitely follow arbitrage strategy. In the next section we will introduce alternative strategies that one can choose from when arbitrage opportunities are not possible.

5.2 Alternative Strategies

Since we are not guaranteed to benefit from arbitrage opportunities in betting market, we will introduce potential strategies that could hopefully yield positive return over betting in matches in Premier League in season 2018-2019. At this stage, what we can do is to come up with certain strategies that anyone could apply to make profit or rely on the models we introduced in Methodology section and compare the results. The simple strategy that one could apply would be to bet always on home team win in all matches played in season 2018-2019. Assuming that the gambler bets 1 euro for each match, he would invest 380 euros in total and the percentage return could be represented as follows:

$$PercentageReturn = \frac{\sum_{i=1}^{380} odd_{homewin} - 380}{380} 100$$

Similarly, we could develop another simple strategy for gambler to bet. Again, assuming that the gambler bets 1 euro for each match, he would invest 380 euros in total for away team win and the percentage return could be represented as follows:

$$PercentageReturn = \frac{\sum_{i=1}^{380} odd_{awaywin} - 380}{380} 100$$

Similarly, we could develop another simple strategy for gambler to bet. Again, assuming that the gambler bets 1 euro for each match, he would invest 380 euros in total for the draw outcome and the percentage return could be represented as follows:

$$PercentageReturn = \frac{\sum_{i=1}^{380} odd_{draw} - 380}{380} 100$$

Besides these strategies, we will also now rely on the models we introduced in **Chapter 3** and bet on the option that is most probable under those models. Meanwhile, in

our case we will be betting on either of these three options: Home Team Win; Draw; Away Team Win. The percentage return formulas will be similar to those before but instead, we will bet on options suggested by Poisson and Dixon-Coles models in separate and then compare these five betting strategies to observe whether any of them results in positive return for over 2018-2019 season.

Chapter 6

Model

In this chapter we will talk about the models we built when giving predictions for match outcomes. In a given match there are only three possible outcomes: home team win, draw and away team win. If number of goals scored by home team outweighs the number of goals scored by away team then home team is considered as winner of the match, however, if the away team scores more goals than home team then away team is considered as winner, otherwise the result is draw.

And referring back to our models we introduced in **Chapter 3** we will predict the expected number of goals each team to score against each other and then insert those expected number of goal values(λ and μ for home and away teams respectively) into equation (3.17) and (3.18) in Poisson and Dixon-Coles models respectively to extract probabilities of potential match score outcomes. And then we will add up the probabilities of the scores where home team scores more than away team and call this sum as win probability for home team. Similarly, we will add up the probabilities of the scores where away team scores more goals than home team and call this sum win probability for away team. And finally the probabilities over the score where home team and away team score same number of goals will add up to draw probability.

In **section 6.1** we will conduct preliminary statistics over the data we gathered for English Premier League through season 2010-2011 up to season 2017-2018. And then in **section 6.2** we will introduce model results for Poisson regression and Dixon-Coles models and compare the results.

6.1 Preliminary Data Analysis

As can be seen from the **Figure 6.1** below, one could observe that Poisson distribution fits both the number of goals scored by home team and away team quite well. Poisson regression, being one of the widely applied regression method over count data, is therefore thought to be good fit for our data, where counts in our case are the goals. However, one should note that when making Poisson regression, we will not treat goals scored by home and away teams differently, we will rather gather all goals scored both by home and away teams under one 'goals column' when implementing Poisson regression and Dixon-Coles method and treat all goals as independently distributed from one another.

$$\lambda = 1.55 \quad \mu = 1.18$$

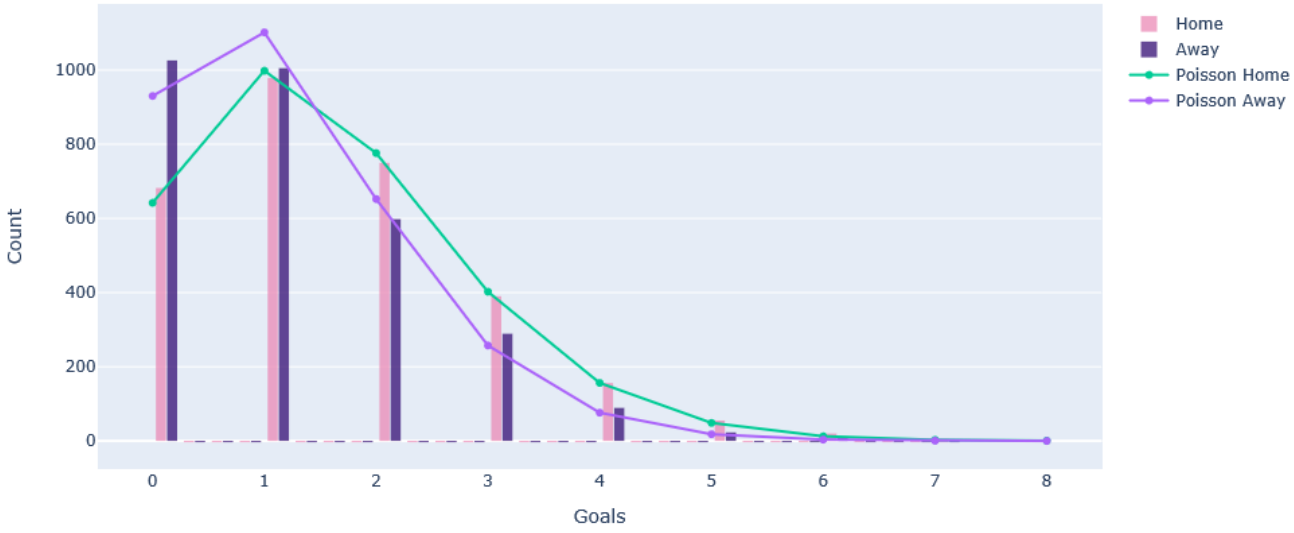


Figure 6.1: Frequency of Observed and Predicted number of home and away goals scored from season 2010-2011 up to season 2017-2018 in Premier League. Pink bars show the frequency of goal values scored by home teams, whilst purple bars demonstrate the frequency of goal values scored by away teams. In addition, green line depicts the predicted number of goals to be scored by home team and purple line shows the predicted number of goals to be scored by away team under the Poisson distribution.

Regarding goal interval in **Figure 6.1**, we should say that it is not decided randomly. The maximum goal scored from season 2010-2011 up to season 2018-2019 is 8, whereas minimum number of goal was 0, which is why our Goals values range from 0 to 8 in the histogram above.

6.1.1 Explanatory Variables in the Generalized Linear Model

Since we already know that Poisson distribution is part of the exponential family of distributions, we could fit a generalized linear model for data that has response variable Poisson distributed. As given in the methodology section, we know that :

$$g(E[\,Y_i]) = x_i^T \beta$$

Our main aim is to find estimates for β vector and since x_i^T stands for the values of explanatory variables for the observation indexed with i, it is β vector that will help one determine the expected value for response variable or $E[\,Y_i]$, which in our case is the average number of goals scored.

All being said, our main goal in this section is to describe the explanatory variables we will be using when predicting number of goals each team will score and furthermore, give rationale for our choice of those explanatory variables.

Initially, we will want our model to reflect teams' strengths when giving goal predictions and difference in teams' strengths exist since not every team is equally competent. Those people following football matches are aware of the notion that certain teams outplay rest of the teams in most of the matches, indeed there are exceptions and sensations occurring in a football match as well as anything else that has randomness in its nature. However, we first, should convince ourselves that all teams are not performing equally well. And in order to visualize the fact that there exists a difference in the quality of teams, at least those competing in Premier League, we will refer to the following Figure that includes teams which took part in Premier League from the season 2010-2011 up to the season 2017-2018:

As can be seen from the **Figure 6.2**, certain teams such as Manchester City are more inclined to score than concede a goal on average. However, teams such as Cardiff have conceded more goals than they scored on average. Such a characteristics should be included in our model so that the predictions we make reflect the fact that some teams are likely to score more than they are to conceded, and some are vice versa.

Another very important factor to take into account when modelling goals would be

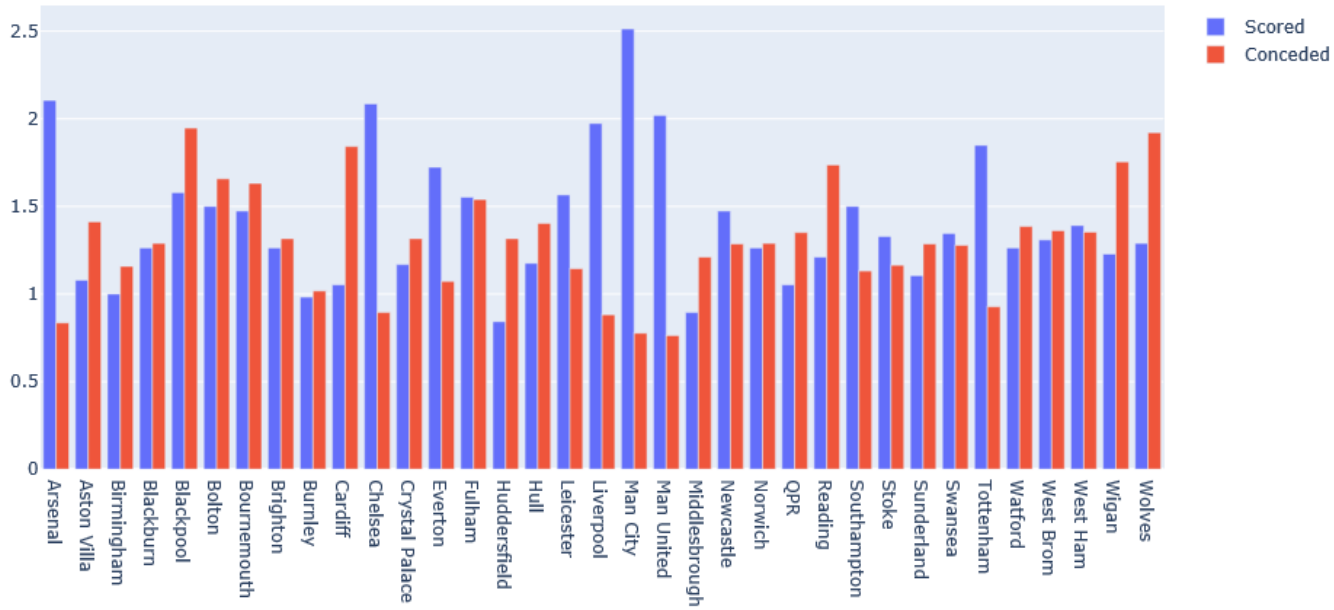


Figure 6.2: Average number of goals scored(blue) and conceded(red) per match by the teams that participated in Premier League from 2010-2011 to 2017-2018.

home advantage effect. There is widely believed phenomenon that if a team plays at its own stadium then that particular team, holding everything else constant, is more advantageous to the team that it is playing against. We will try to observe whether such a phenomenon exists in our data as well:

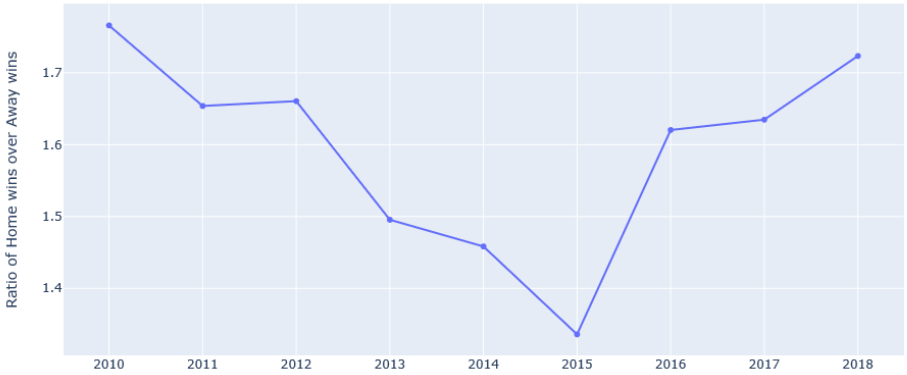


Figure 6.3: Ratio of games won home to games won away for all teams participating in English Premier League from season 2010-2011 up to season 2017-2018.

Without rigorously proving that a team playing at its own stadium is more advantageous to win a certain match than that particular team playing the same match away, referring to **Figure 6.3** one could easily observe that number of games won home is more than that of games won away during the time span of our selected data.

To put in a nutshell, we have gained useful insights regarding the inclusion of explanatory variables into our model. Due to the existing discrepancy between teams when it comes to scoring and conceding goals, we will add attacking and defensive strengths of teams, which are purely extracted through regression output on the basis of number of goals each team score and concede, alongside with home advantage factor.

6.2 Model Outputs

6.2.1 Poisson Model

After modelling Poisson distributed random variable Y, goals in our case, by using Team, Opponent and home advantage variables, we will try to use (3.17) to predict score probability predictions for matches between teams performing in Premier League season 2018-2019. As we noted earlier in **section 3.7**:

$$U_{i,j} \sim Poisson(\alpha_i \beta_j \gamma)$$

$$V_{i,j} \sim Poisson(\alpha_j \beta_i)$$

$for i, j = 1, 2, ...m, i \neq j$ and we already explained what m stands for in **section 4.1**. $U_{i,j}$ stands for the number of goals team i scores against team j at home and $V_{i,j}$ is the number of goals team i scores against team j away. Moreover, λ is expected value of $U_{i,j}$, which is represented as $\alpha_i \beta_j \gamma$ and μ is the expected value $V_{i,j}$, denoted as $\alpha_j \beta_i$. And as discussed earlier, $\alpha_i, \beta_j, \gamma > 0$. The subscripts i and j stand for home and away teams, and α and β values representing attack and defensive strength respectively, while γ parameter denotes home advantage coefficient. After obtaining regression outputs we will show how λ and μ parameters, denoting expected number of goals home and away team to score respectively, are obtained. But before, let's describe the coefficients we will obtain from regression output:

$$\tilde{\alpha_i} = log\alpha_i \quad \tilde{\alpha_j} = log\alpha_i \quad \tilde{\beta_i} = log\beta_i \quad \tilde{\beta_j} = log\beta_j \quad \tilde{\gamma} = log\gamma$$

Since we are using logarithmic link function:

$$\lambda_{ij} = \alpha_i \beta_j \gamma = e^{\tilde{\alpha_i} + \tilde{\beta_j} + \tilde{\gamma}}$$

$$\mu_{ji} = \alpha_j \beta_i = e^{\tilde{\alpha_j} + \tilde{\beta_i}}$$

$i, j = 1, ...m$ and $j \neq i$.

Meanwhile, we have sampled training data so that all teams present in test data have already performed at least one season in our training data. In other words, in order to give predictions for teams participating in Premier League in season 2018-2019, we should have all those teams in our training data. This step is necessary since each year three teams relegate to second division and some different three teams from second division gains the right to participate in Premier League.

Now, regarding the equations above, one should note that referring to Poisson regression output in Appendix 1, $\tilde{\alpha}$ coefficients for teams are differentiated from $\tilde{\beta}$ and $\tilde{\gamma}$ coefficients by addition of 'team' word before the actual name of the team. Whereas $\tilde{\beta}$ coefficients for teams are represented by addition of 'opponent' word before the actual name of the team, whereas $\tilde{\gamma}$ coefficient, denoting home advantage is represented as 'home' in the model output. And calculation of $\lambda_{i,j}$ values can be represented as follows:

$$\lambda_{ij} = \alpha_i \beta_j \gamma = e^{\tilde{\alpha_i} + \tilde{\beta_j} + \tilde{\gamma}} = e^{\mathbf{X'B}}$$

where

$$\mathbf{X'} = \begin{bmatrix} \mathbf{x'_i} \\ . \\ \mathbf{x'_m} \\ . \\ \mathbf{x'_{m+j}} \\ . \\ \mathbf{x'_{2m}} \\ . \\ \mathbf{x'_{2m+1}} \end{bmatrix}, \mathbf{B} = \begin{bmatrix} \tilde{\alpha_i} \\ . \\ \tilde{\alpha_m} \\ . \\ \tilde{\beta_j} \\ . \\ \tilde{\beta_m} \\ . \\ \tilde{\gamma} \end{bmatrix}$$

$i, j = 1, ...m$ and $j \neq i$.

Where $\mathbf{B}$ is the parameter vector as in Appendix 1 but for m teams in season 2018-2019 rather than n teams in our training data. And in our code we have created $\mathbf{X'}$ being the vector of 2m+1 elements, and we have m teams(not n, since we are predicting results for m teams in test data) repeating twice in $\mathbf{X'}$ vector for the simplicity in calculation of $\lambda_{i,j}$ and $\mu_{j,i}$ values, whilst (2m+1)th, or the last element in $\mathbf{X'}$ vector is 1 in case of calculation of $\lambda_{i,j}$ and 0 in case of calculation of $\mu_{j,i}$ values.Meanwhile since we are repeating teams in $\mathbf{X'}$ vector, ith and (m+i)th elements are the same for $i = 1, ...m$. In calculation of $\lambda_{i,j}$, except ith, (m+j)th and (2m+1)th(last) elements, rest of the elements in the $\mathbf{X'}$ vector are zero and those with ith and (m+j)th elements are 1. ith element of $\mathbf{X'}$ vector is multiplied by the ith element of $\mathbf{B}$ vector, or $\tilde{\alpha_i}$ and (n+j)th element of $\mathbf{X'}$ vector is multiplied by the (m+j)th element of $\mathbf{B}$ vector, or $\tilde{\beta_j}$. And finally, (2m+1)th element of $\mathbf{X'}$ vector, or 1 is multiplied by the last element of $\mathbf{B}$ vector, or $\tilde{\gamma}$.Similarly, one can show the calculation of $\mu_{j,i}$ values in the following way:

$$\mu_{ji} = \alpha_j \beta_i = e^{\tilde{\alpha_j} + \tilde{\beta_i}} = e^{\mathbf{X'B}}$$

where

$$\mathbf{X'} = \begin{bmatrix} \mathbf{x'_j} \\ . \\ \mathbf{x'_m} \\ . \\ \mathbf{x'_{m+i}} \\ . \\ \mathbf{x'_{2m}} \\ . \\ \mathbf{x'_{2m+1}} \end{bmatrix}, \mathbf{B} = \begin{bmatrix} \tilde{\alpha_j} \\ . \\ \tilde{\alpha_m} \\ . \\ \tilde{\beta_i} \\ . \\ \tilde{\beta_m} \\ . \\ \tilde{\gamma} \end{bmatrix}$$

$i, j = 1, ...m$ and $j \neq i$.

In calculation of $\mu_{j,i}$, except jth and (m+i)th elements, rest of the elements in the $\mathbf{X'}$ vector are zero and those with jth and (m+i)th elements are 1. jth element of $\mathbf{X'}$

vector is multiplied by the jth element of $\mathbf{B}$ vector, or $\tilde{\alpha_j}$ and (m+i)th element of $\mathbf{X'}$ vector is multiplied by the (m+i)th element of $\mathbf{B}$ vector, or $\tilde{\beta}_i$. And finally, (2m+1)th element of $\mathbf{X'}$ vector, or 0 is multiplied by the last element of $\mathbf{B}$ vector, or $\tilde{\gamma}$.

Since elements of our explanatory variables- **Team** and **Opponent**, are actual team names, we have observed Arsenal to be the intercept term. In Appendix 1, one could realize that $\tilde{\alpha}_{Arsenal}$ coefficient is given, however, $\tilde{\beta}_{Arsenal}$ coefficient is set to zero. Other than Arsenal, $\tilde{\beta}$ coefficients for all teams are given. Regarding interpretation of $\tilde{\alpha}$ parameters, for example, for the team Arsenal $e^{\tilde{\alpha}_{Arsenal}}$ or $\alpha_{Arsenal}$(attacking strength) stands for the average number of times Arsenal would score more goals than the overall average number of goals scored by all teams, holistically speaking. When it comes to interpretation of $\tilde{\beta}$ parameters, again giving the example of Arsenal, $\tilde{\beta}_{Arsenal}$, which is zero and thus making $e^{\tilde{\beta}_{Arsenal}}$ or $\beta_{Arsenal}$(defensive strength) be 1, stands for the average number of times Arsenal would concede more goals than the overall average number of goals conceded by all teams, again holistically speaking. In case of Arsenal, it is held that Arsenal would have exactly defensive strength of 1, so that defensive strength parameters for the rest of the teams would be multiplied with that of Arsenal, or be multiplied by 1. The parameter γ indicates the average number of times a home team would score more goal than an away team. Further more, $\lambda_{i,j}$ stands for the expected number of goals team indexed i would score against team indexed j, while $\mu_{j,i}$ is the expected number of goals team indexed j would score against team indexed i.

Once we have λ_{ij} and μ_{ji} obtained from regression output for $i, j = 1, ...m$ and $j \neq i$, we will then insert these values to the following equation below to get possible match score probabilities between team i and j, again where $i, j = 1, ...m$ and $j \neq i$.

$$P(U_{i,j} = u; V_{i,j} = v) = \frac{\lambda_{ij}^{u} e^{-\lambda_{ij}}}{u!} \frac{\mu_{ji}^{v} e^{-\mu_{ji}}}{v!}, \tag{6.1}$$

for i,j=1,...m and j$\neq i$

The above equation is different from that in (3.17) in a sense that we have indexed λ and μ values. Those indexes exist in (6.1) due to the fact that in this very specific case, each team has varying expected number of goals to score depending on the team

they are playing with and home advantage. After we predict the probabilities for each potential outcome of the match played between teams, we will assign the probability for home team to win, for away team to win and finally for draw. For example, calculation of probability for home team to win, again assuming that team i plays at home and scores u goals whereas team j plays away and scores v goals, for each teams playing home would be as follows:

$$P(HomeTeamWin) = \sum_{u=1}^{maxgoals} \sum_{v=0}^{u-1} \frac{\lambda_{ij}^{u} e^{-\lambda_{ij}}}{u!} \frac{\mu_{ji}^{v} e^{-\mu_{ji}}}{v!}, \tag{6.2}$$

for i,j=1,...m and j$\neq i$; u,v$\in N$.

Similarly, calculation of draw probability for each match could be described as follows:

$$P(Draw) = \sum_{u=v=0}^{maxgoals} \frac{\lambda_{ij}^{u} e^{-\lambda_{ij}}}{u!} \frac{\mu_{ji}^{v} e^{-\mu_{ji}}}{v!}, \tag{6.3}$$

for i,j=1,...m and j$\neq i$; u,v$\in N$.

And finally, calculation of probability of away team win for each match can be described as follows:

$$P(AwayTeamWin) = \sum_{v=1}^{maxgoals} \sum_{u=0}^{v-1} \frac{\lambda_{ij}^{u} e^{-\lambda_{ij}}}{u!} \frac{\mu_{ji}^{v} e^{-\mu_{ji}}}{v!}, \tag{6.4}$$

for i,j=1,...m and j$\neq i$; u,v$\in N$.

Before we move on any further, as noted in **section 4.1** it should be made clear that our training data contains all match results from season 2010-2011 up to season 2017-2018 in Premier League. And after building models over the training data, we will be applying our model to matches played in 2018-2019 season in Premier League. In other words, our test data contains matches from season 2018-2019.

And as we already noted earlier in **section 6.1** earlier, the maximum number of goals is taken to be 8. Once we have the probability for all home teams to win, lose and take draw in matches using (6.2), (6.3) and (6.4) in season 2018-2019 obtained, we will assign one of those labels(home/draw/away) for every game with the label that has highest probability of occurrence. And we will call it predicted result of the game under

Poisson Regression. And referring to **Figure 6.4**, one could be familiarized with the confusion matrix we built. Confusion matrix is a table layout presenting performance of an algorithm. As can be seen from the figure below, our accuracy rate of correctly predicting the final match outcomes for Premier League during season 2018-2019, using 2010-2011 up to 2017-2018 match outcome data, is approximately 58.42 percent which obviously exceeds a random guess that has one third of success rate for three possible match outcomes. Since we have our confusion matrix ready, now we could interpret the

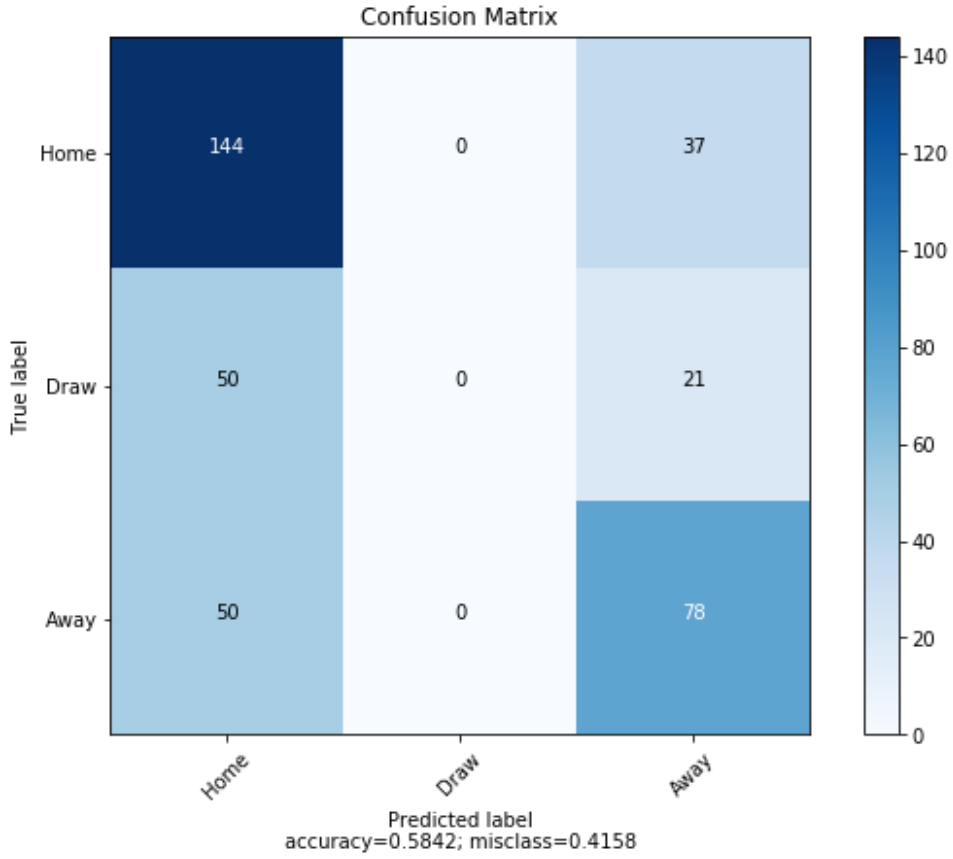


Figure 6.4: Confusion matrix for predicted and actual result of the games under the Poisson model.

performance of the Poisson model. On the left, we have True label, denoting actual results. And below we have Predicted match outcome for home team to win, draw and away team to win. Summing diagonal line from upper left to lower right will yield in the number of matches which we correctly predicted, the rest are misclassified instances.

The thing that should catch one's interest is Poisson model's very poor performance in predicting Draw outcomes. In fact, one could observe that Poisson model predicts

no Draw outcome for any of those 380 matches, where predicted outcome is simply the event(home/draw/away) that has highest probability of occurrence. On the way to investigate factors that might be causing such deficiencies of Poisson model, we present a matrix of match results in **Figure 6.5**, where the difference between average of Poisson predicted match score probabilities and actual percentage of matches ended with those scores from 2010-2011 to 2017-2018 is presented.

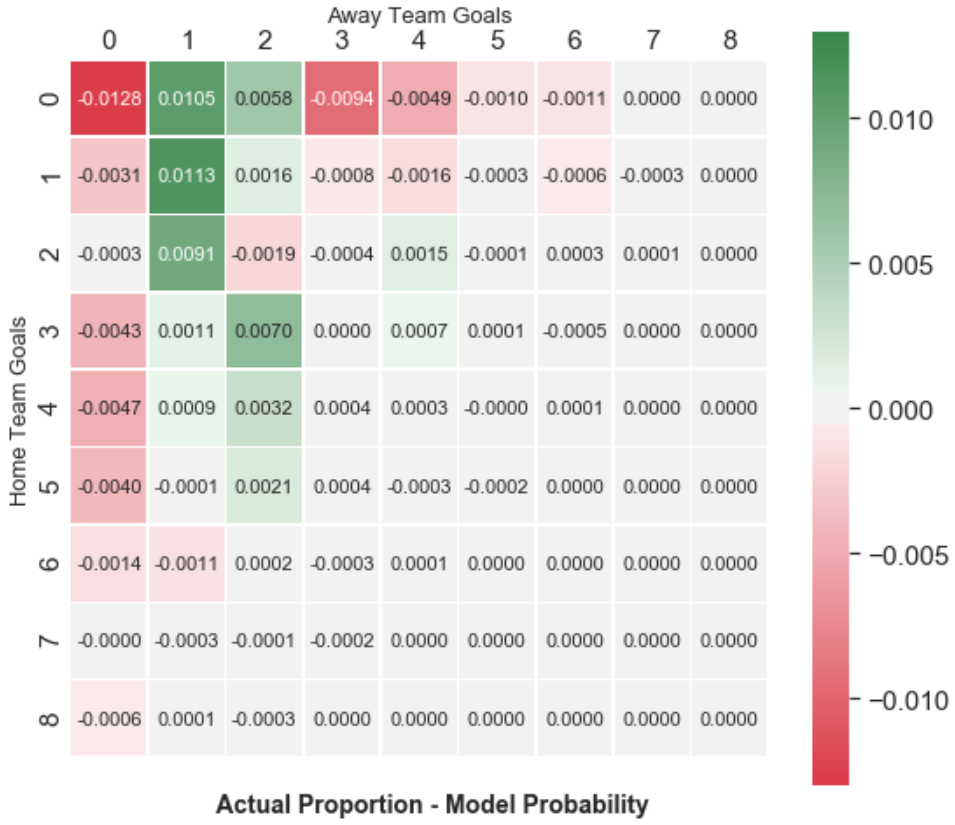


Figure 6.5: Difference between average of Poisson predicted score probabilities and percentage of matches that ended with given results from season 2010-2011 up to season 2017-2018 in Premier League.

.

In **Figure 6.5**, one should note that when the colour strength demonstrates the level of disagreement. The greener the cells become, the less weight Poisson model puts on those scores. On the contrary, the redder the cells become, the more weights Poisson model assigns for those scores. And it is easily observed that both green and red colours get most darker in the northwest square of the matrix where we have match scores being 0 and 1. Obviously, Poisson model is having hard time predicting the match score out-

come that has combination of 0 and 1 in it. With the hope of solving this problem, we will refer to the other model we introduced in **Chapter 3**, Dixon-Coles model.

6.2.2 Dixon-Coles Model

In previous section we realized that Poisson model is not very good at predicting Draw results, while referring to **Figure 6.4** we have 71 of 380 matches ended in Draw in Premier League. Apart from that, referring to **Figure 6.5** we realized that Poisson model seems to be inferior model when giving predictions for matches that could end with score outcome that has combination of 0 and 1 in it. Once we numerically maximize (3.22) we will obtain coefficients for parameters that we discussed in **subsection 3.7.2**.Estimated parameter coefficients for Dixon-Coles model will be given in Appendix 2. First $m(20)$ coefficient estimates could be thought as $\tilde{\alpha}$ values and next $m(20)$ coefficients could be thought as $\tilde{\beta}$ values and last one as $\tilde{\gamma}$ value as in previous section. Since Dixon-Coles model is built by adjusting Poisson model, we will apply exactly the same procedure as in **subsection 6.2.1** till (6.4). however, instead of (6.1), (6.2), (6.3) and (6.4), we will use the following equations adjusted for Dixon-Coles model:

$$P(U_{i,j} = u;\, V_{i,j} = v) = \tau_{\lambda_{i,j},\mu_{i,j}}(u_{i,j}, v_{i,j}) \frac{\lambda_{ij}^{u} e^{-\lambda_{ij}}}{u!} \frac{\mu_{ji}^{v} e^{-\mu_{ji}}}{v!}, \tag{6.5}$$

for i,j=1,...m and j$\neq i$

$$P(HomeTeamWin) = \sum_{x=1}^{maxgoals} \sum_{v=0}^{u-1} \tau_{\lambda_{i,j},\mu_{i,j}}(u_{i,j}, v_{i,j}) \frac{\lambda_{ij}^{u} e^{-\lambda_{ij}}}{u!} \frac{\mu_{ji}^{v} e^{-\mu_{ji}}}{v!}, \tag{6.6}$$

for i,j=1,...m and j$\neq i$; u,v$\in N$.

Similarly, calculation of draw probability for each match could be described as follows:

$$P(Draw) = \sum_{u=v=0}^{maxgoals} \tau_{\lambda_{i,j},\mu_{i,j}}(u_{i,j}, v_{i,j}) \frac{\lambda_{ij}^{u} e^{-\lambda_{ij}}}{u!} \frac{\mu_{ji}^{v} e^{-\mu_{ji}}}{v!}, \tag{6.7}$$

for i,j=1,...m and j$\neq i$; u,v$\in N$.

And finally, calculation of probability of away team win for each match can be described as follows:

$$P(AwayTeamWin) = \sum_{v=1}^{maxgoals} \sum_{u=0}^{v-1} \tau_{\lambda_{i,j},\mu_{i,j}}(u_{i,j}, v_{i,j}) \frac{\lambda_{ij}^{u} e^{-\lambda_{ij}}}{u!} \frac{\mu_{ji}^{v} e^{-\mu_{ji}}}{v!}, \quad (6.8)$$

for i,j=1,...m and j$\neq i$; u,v$\in N$.

Once we have the probability for all home teams to win, lose and take draw in matches using (6.6), (6.7) and (6.8) in season 2018-2019 obtained, we will assign one of those labels(home/draw/away) for every game with the label that has highest probability of occurrence. And we will call it predicted result of the game under Dixon-Coles model. Since we had assigned more weights over probabilities for scores ending with $0-0$ and $1-1$, we would expect Poisson model prediction and Dixon-Coles model prediction be different for low scoring games. In order to observe this phenomenon in practice, we shall have a look at the difference in predicted score probabilities between Poisson and Dixon-Coles model for a random match(Arsenal-Bournemouth in our case).

As could be seen from the **Figure 6.6**, upper-left part of the matrix, which is the result of adjustment we made using τ function in **subsection 3.7.2**. Obviously, most of the difference in predicted probabilities for potential match score result between Arsenal and Bournemouth is observed for the scores that has combination of 0 and 1 simultaneously in it, which is to be expected thanks to τ function.

And referring to **Figure 6.7**, one could be familiarized with the confusion matrix we built. As can be seen from the figure below, our accuracy rate of correctly predicting the final match outcomes for Premier League during season 2018-2019, using 2010-2011 up to 2017-2018 match outcome data, is approximately 58.16 percent which obviously exceeds a random guess that has one third of success rate for three possible match outcomes, however, performance of Dixon-Coles model is falling behind the performance of Poisson model. We can try to understand the main rationale behind that by relying on the figure below:

As could be seen from the **Figure 6.6**, the only difference in performance level

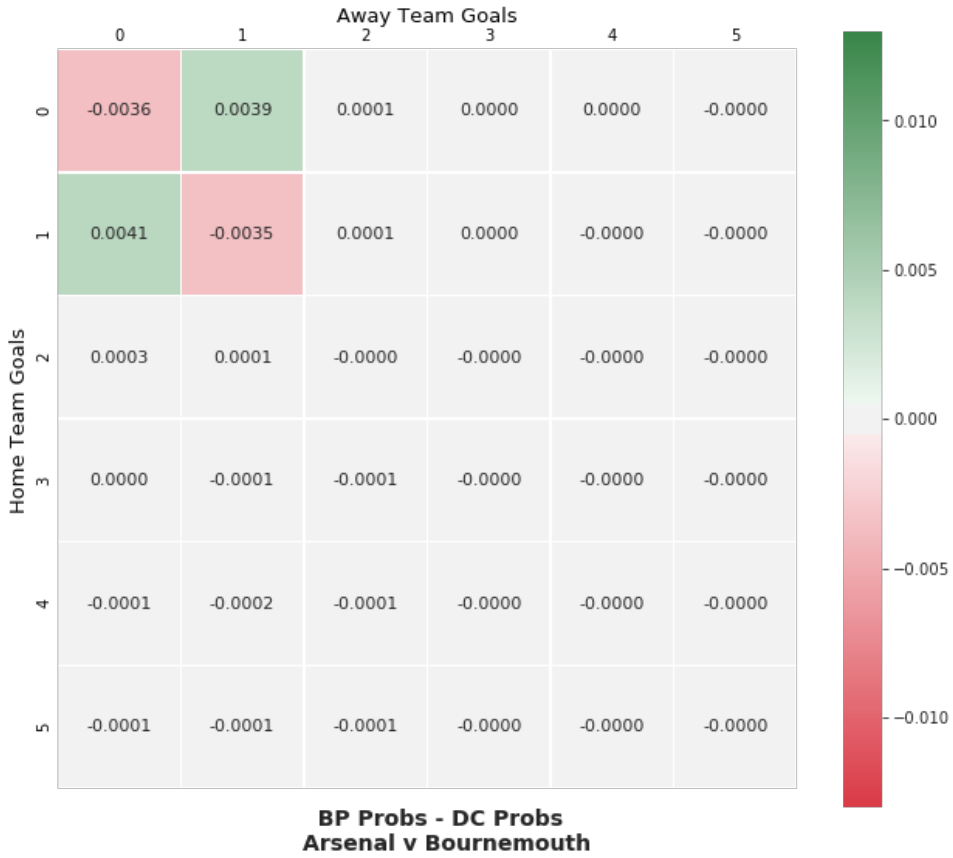


Figure 6.6: Difference in predicted score probabilities between Poisson and Dixon-Coles model for a match between Arsenal and Bournemouth.

between Poisson and Dixon-Coles models in terms of match outcome prediction is 2 wrongly predicted Draws by Dixon-Coles model. Dixon-Coles model assigns more weight over probability of match outcomes for Draw results- $0-0$ and $1-1$. However, in our case it seems that Dixon-Coles model is not solving the problem we are having, on the contrary, it is further decreasing the predictive accuracy of the model. At the very least, for our data, we realize that Poisson model outperforms Dixon-Coles model in terms of accuracy of model performance. However, this need not always necessarily be the case. Although being outperformed by Poisson model in our data, Dixon-Coles model will still be referred to when observing possibility of positive return over betting.

Yet, before we move on to start talking about results of betting strategies and their corresponding returns, we will try to look at performance of both models from different angle. Going back, in equations (6.2), (6.3) and (6.4) we obtain the probability of home team win, draw, and away team win for each match and when assigning label for

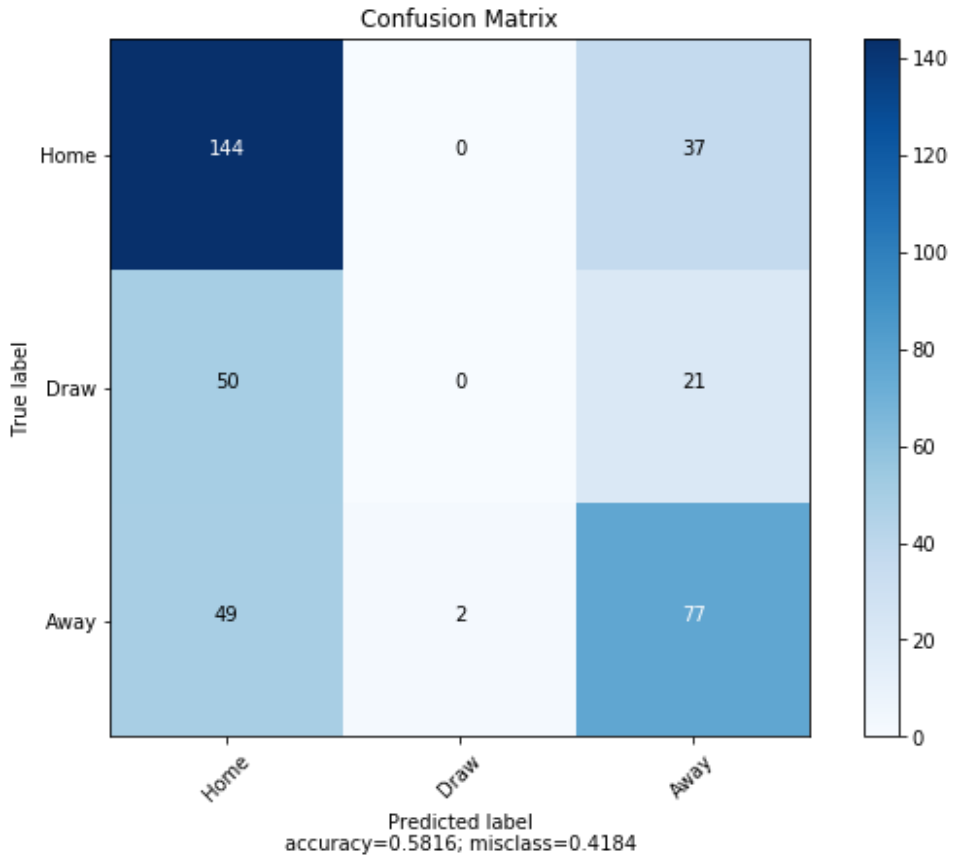


Figure 6.7: Confusion matrix for predicted and actual result of the games under the Dixon-Coles model.

match outcome we predicted the outcome for the match result that had highest probability of occurrence. Now, we will create bins with fixed probability intervals and then compare our model accuracy per each bin to see the performance of model favourable prediction(outcome that has the highest probability of occurrence).

As can be seen from the **Table 6.1**, it is the interval for predicted probability for the favourable outcome ranging between 70-80 percent, where the predicted outcome is realized 86 percent of the time. On the other hand, once the probability for the model favourable outcome is in interval ranging from 30 to 40 percent, we have only 41.67 percent of the predictions realized. We could extract the same table for Dixon-Coles model.

Again, referring to **Table 6.2**, it is the interval for predicted probability for the favourable outcome ranging between 70-80 percent, where the predicted outcome is realized 87.8 percent of the time. On the other hand, once the probability for the model

Table 6.1: Accuracy of Model favorable prediction per probability interval under Poisson Model

Model Predicted Probability	Number of matches	Number of matches won by model favorite	Percentage of matches won by model favorite
30%-40%	48	20	41.67%
40%-50%	113	52	46.02%
50%-60%	102	60	58.82%
60%-70%	60	43	71.67%
70%-80%	43	37	86.05%
80%-90%	14	10	71.43%
90%-100%	0	0	None
Total	380	222	58.42%

Table 6.2: Accuracy of Model favorable prediction per probability interval under Dixon-Coles Model

Model Predicted Probability	Number of matches	Number of matches won by model favorite	Percentage of matches won by model favorite
30%-40%	55	23	41.82%
40%-50%	112	51	45.54%
50%-60%	100	60	60.00%
60%-70%	59	42	71.19%
70%-80%	41	36	87.80%
80%-90%	13	9	69.23%
90%-100%	0	0	None
Total	380	221	58.16%

favourable outcome is in interval ranging from 30 to 40 percent, we have only 41.82 percent of the predictions realized.

Now that we have discussed performance of both models, it would be interesting to see if these models can be of any use for an enthusiastic gambler. With that being said, we are proceeding with next chapter to see the potential benefits of such models.

Chapter 7

Results

In this chapter we will be talking about what additional information previous models we built could give for someone gambling in betting market. In section **section 7.1** we will be talking about teams' performance under goals models(both Dixon-Coles and Poisson models are referred as goals models in literature). And later, we will try to observe if relying on these models could make one richer than otherwise.

7.1 Team ranking

In previous chapter we talked about accuracy of models, confusion matrix and other descriptive statistics methods measuring performance of our models. In this section however, we will see whether those models speak anything about the team rankings in Premier League in season 2018-2019. In other words, since we relied on training data, matches' data from season 2010-2011 to season 2017-2018 included, we want to know if knowing the simple match results from 2010-2011 to 2017-2018 could take one beyond the prediction for matches over test data. That being said, we will now investigate whether Dixon-Coles model could predict the teams' standpoint in Premier League season 2018-2019.

The Plot below depicts the teams' stances purely by their attacking and defensive strengths that we obtained through Dixon-Coles model. Before we start interpreting the plot above, we should be able to understand the axis values. Regarding y-axis, we have attacking strength value represented, which we explained in Chapter 6. And the x-axis indicates defensive strengths of teams performing in Premier League in season

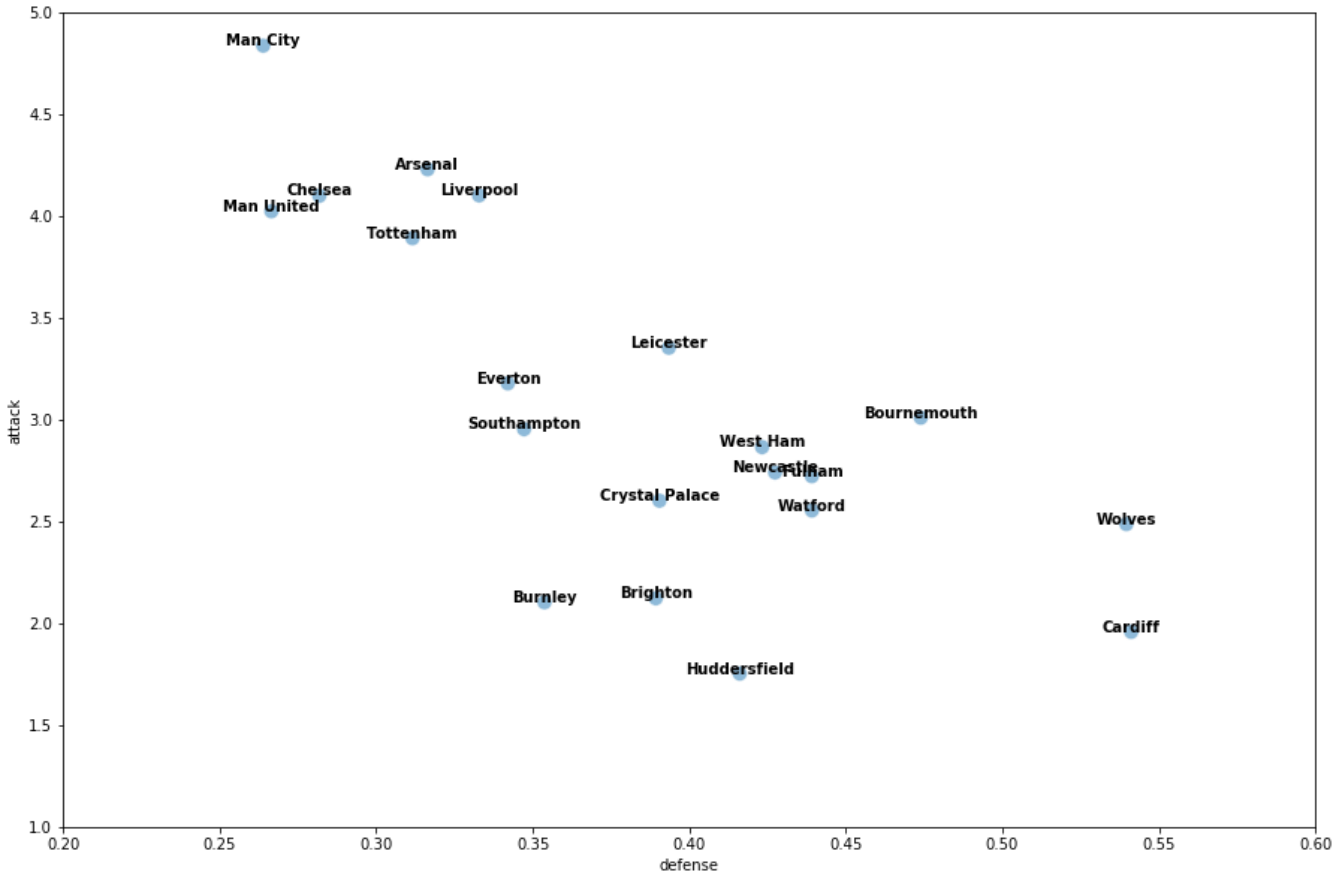


Figure 7.1: Teams' standpoint in terms of their attacking and defensive strengths.

2018-2019. The higher the attacking strength a team has, the more likely that team is to score goals against others. On the contrary, the lower the defensive strength a team has, the less likely that team will concede a goal. Having those said, one could see that Manchester City is the team performing best among all regarding both attacking and defensive strengths, whereas, Cardiff and Huddersfield seem to be worst performing teams in Premier League season 2018-2019 according to our predictions. In general, teams being located on the upper left part of the Figure 7.1 are supposed to be overperforming the rest, whereas teams being located on the lower right of the Figure 7.1 are supposed to be worst performing teams. Interestingly, Manchester City, locating on the most upper left of our plot, according to Dixon-Coles model predictions actually became champion in Premier League season 2018-2019. What's more, the six teams locating on the most upper left part of Figure 7.1 took first six places in the Premier League in season 2018-2019, whilst Huddersfield and Cardiff were two out of three teams to be relegated to the second

division. Apparently, attacking and defensive strengths of teams say a lot regarding team performance. However, what one might be interested in is the very reason of Manchester City's overperforming others especially due to the fact that as many football fans are familiar with it, Manchester City was not the amongst the top performers in Premier League early 2000s. It was teams such as Chelsea, Manchester United, Liverpool etc. which were the main players not only in Premier League but also in European Championships. In order to understand why say model does not predict Manchester United be the champion for the season 2018-2019, but it does predict Manchester City be champion, we will try to analyze the ratio of average goals both teams scored and conceded to the average number of goals scored and conceded in Premier League across years 2010 to 2018. The reason why we do this is both Poisson and Dixon-Coles models are considered to be goals models since the number of goals teams score and concede play pivotal role in making the predictions:

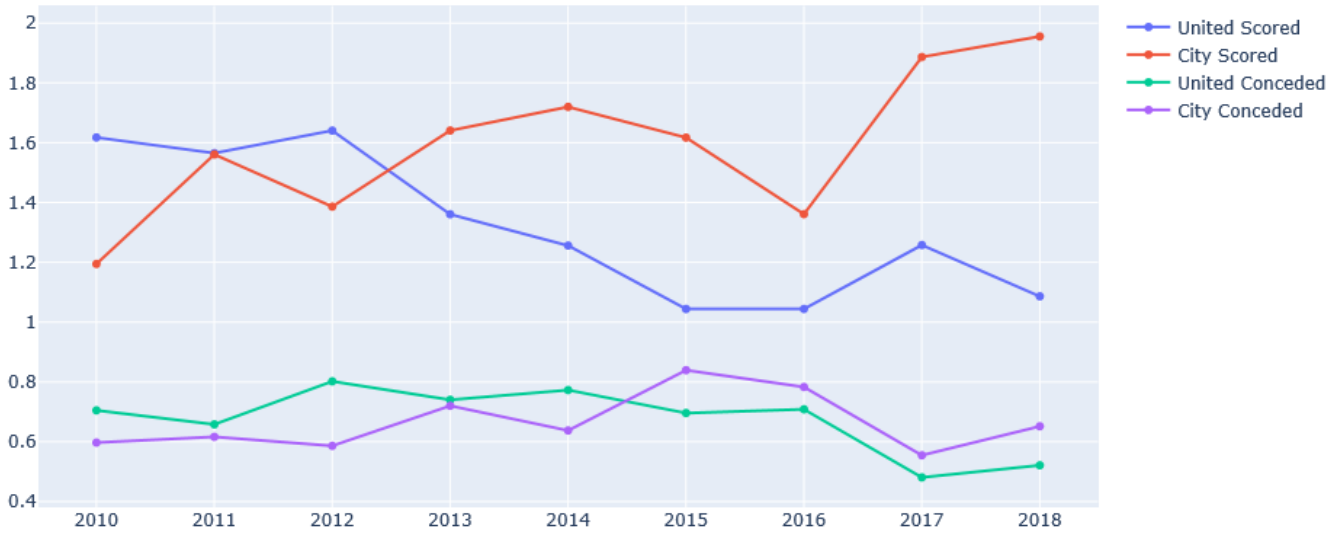


Figure 7.2: Teams' standpoint in terms of their attacking and defensive strengths.

Now, the red line indicates the ratio of average number of goals Manchester City scored to the average number of goals scored by all teams from 2010-2018, whilst the blue line shows the same for Manchester United. And the purple line demonstrates the ratio of average number of goals conceded by Manchester City to the average number of goals conceded by all teams from 2010-2018, while green line shows the same for Manchester

United. As can be seen from the figure above, it is easily realized that Manchester City started to outperform Manchester United starting from season 2012-2013 when it comes to ratio of average number of goals scored by those teams to the average number of goals scored by all teams in Premier League. Then it might be understandable that Manchester City scored more goals than Manchester United during seasons starting from 2012-2013 and that is the rationale behind Dixon-Coles model selecting Manchester City as a champion but not Manchester United. Regarding the ratio of average number of goals conceded by both teams, we could realize that it is fluctuating across years between those two teams and not much of a conclusion could be made out of Figure 7.2 except the one that starting from 2015 Manchester City began to conceded more goals than Manchester United.

7.2 Return

Now that we have made our predictions for each 380 game over the season 2018-2019 in Premier League, one would like to test whether those predictions could be financially of any use. In this section, we will simulate 1 euro bet on each match under different betting strategies as discussed in **section 5.2**. Referring to following table, one could see the comparison of returns obtained through various betting strategies:

Table 7.1: Percentage return over betting 1 euro in every match in Premier League season 2018-2019 per strategy.

Strategy	Return
Bet always on home team	2.51%
Bet always on draw	-5.35%
Bet always on away team	-28.32%
Bet on Poisson model favorable outcome	3.30%
Bet on Dixon-Coles model favorable outcome	2.30%

As can be seen from the Table 7.1, it is betting strategy following Poisson model favourable outcome that yields in the highest percentage return. Return we obtained from Dixon-Coles model is quite disappointing taking into account that a simple strategy such as betting always on home team win can outperform Dixon-Coles model in terms

of percentage return. Betting on draw seems to result in lowest percentage return, which may suggest that betting on draw outcome on a consistent basis actually makes one worse off. Similarly, betting on away team win also results in negative percentage return. Having those said, Basic Poisson model, built with match data from season 2010-2011 to 2017-2018, seems to yield in the highest return for Premier League season 2018-2019.

Chapter 8

Conclusion and Future Work

The purpose of this thesis project was to build Poisson and Dixon-Coles models over match data from season 2010-2011 to season 2017-2018, compare those models, research the possibility of realizing positive return over investment in betting market and observe whether such models could outperform simple strategies in terms of percentage return in Premier League season 2018-2019. And we managed to beat the accuracy of a random guess over a direct match outcome by relying on both models we built. Furthermore, by betting on Poisson model favorable outcome, we managed to beat simple strategies in terms of percentage return (3.30%). Although supposed to be improved version of Poisson model, Dixon-Coles was outperformed both in terms of accuracy and percentage return by Poisson model and it failed to be beat home team win strategy again in terms of percentage return.

8.1 Reliability of Results

In our project work we used so-called goals models when predicting the direct match outcome(win-draw-lose). In a sense, modelling goals and then inserting expected number of goals for each team into probability mass functions by introducing independence of home and away goals to later obtain probability of each outcome is an indirect method of obtaining predictions for direct match results. In our case, since Dixon-Coles model was outperformed by Poisson model, one might be interested in deciphering potential existing correlation between goals scored by home and away teams except scores with combination of 0 and 1 goals. It could be possible to introduce correlation between goals scored by

home and away teams with different combination of scores and later see whether any improvement is made over Poisson model.

Besides improving models, one could decide to set a probability threshold as his benchmark when betting. For example, in **Table 6.1** and **Table 6.2** we have introduced percentage of matches won by model favorable outcome when probability intervals are fixed. Consequently, one might choose to bet only on those matches where model(Poisson or Dixon-Coles) predicted probability of favourable outcome is between 70-80% interval(since it has the highest accuracy in our case).

8.2 Future Improvements

What might certainly experiment is the inclusion of time factor into parameter estimation over Dixon-Coles model. The need for inclusion of time factor is derived from the very fundamental phenomenon that in football teams do not show consistent performance. For example, as discussed in **section 7.1**, the Dixon-Coles model selects Manchester City as the most favorable candidate for championship over season 2018-2019. However, in case Manchester City goes through financial or managerial issues, this could reflect on team performance for the following seasons. However, goals models do not take time effect into consideration when estimating parameters.

References

[1] M. J. Moroney. FACTS from figures, 3rd edn.. Penguin: London, 1956. pages 16

[2] C. Reep. Skill AND CHANCE in BALL GAMES. Journal of the Royal Statistical Society Series A 131: 581-585, 1971. pages 16

[3] M. J. Maher. Modelling ASSOCIATION FOOTBALL scores. Statistica Neerlandica, 1982. pages 16

[4] M.J. Dixon, S.C. Coles. Modelling ASSOCIATION FOOTBALL scores AND inefficiencies in the FOOTBALL betting MARKET. Applied Statistics, 1997. pages 17

[5] H. Rue, O. Salvesen. Prediction AND retrospective ANALYSIS of soccer MATCHES in A LEAGUE. Statistician, 2000. pages 18

[6] M. Crowder, M. Dixon, A. Ledford, M. Robinson. DYNAMIC modelling AND prediction of English FOOTBALL LEAGUE MATCHES for betting. Statistician, 2002. pages 18

[7] D. Forrest, R. Simmons. FORECASTING sport: The BEHAVIOUR AND PERFORMANCE of FOOTBALL tipsters. International Journal of Forecasting, 2000. pages 18

[8] J. Goddard. Regression models for FORECASTING GOALS AND MATCH results in ASSOCIA- tion FOOTBALL. International Journal of Forecasting, 2005. pages 18

[9] B. Hamadani. Predicting the outcome of NFL GAMES using MACHINE LEARNING. Stan- ford University, 2006. pages 18

[10] A. Adam. GENERALISED LINEAR model for FOOTBALL MATCHES prediction. KULeuven, 2016. pages 19

[11] M. Tavakol, H. Zafartavanaelmi and U. Brefeld. FEATURE EXTRACTION AND AGGREGA- tion for Predicting the Euro 2016. Leuphana University of Luneburg, 2016.

pages 19

[12] S. Kampakis, W. Thomas. Using MACHINE LEARNING to Predict the Outcome of English County twenty over Cricket MATCHES. University College London. pages 19

[13] N. Tax, Y. Joustra. Predicting The Dutch FOOTBALL Competition Using Public DATA: A MACHINE LEARNING APPROACH. Transactions on Knowledge and Data Engineer- ing, 2015. pages 19

[14] Annette J Dobson and Adrian G Barnett. An introduction to generalized linear models. Chapman and Hall/CRC, 2008.

[15] Agresti, Alan. (2013). Categorical Data Analysis. Hoboken, New Jersey: John Wiley Sons.

[16] Simeon Denis Poisson. Recherches sur la probabilite des jugements en mati‘ere criminelle et en mati‘ere civile precedees des r‘egles generales du calcul des probabilites par sd poisson. Bachelier, 1837.

[17] Historical football results and betting odds data. https://www. football-data.co.uk/data.php. Accessed: 2020-08-01.

Appendices

Appendix 1

Dep. Variable:	goals	**No. Observations:**	6080
Model:	GLM	**Df Residuals:**	6010
Model Family:	Poisson	**Df Model:**	69
Link Function:	log	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-8779.0
Date:	Sun, 09 Aug 2020	**Deviance:**	6823.0
Time:	21:37:42	**Pearson chi2:**	5.93e+03
No. Iterations:	5		

	coef	**std err**	**z**	**P> |z|**	**[0.025**	**0.975]**
team[Arsenal]	0.2893	0.071	4.090	0.000	0.151	0.428
team[Aston Villa]	-0.3342	0.086	-3.890	0.000	-0.503	-0.166
team[Birmingham]	-0.4132	0.174	-2.377	0.017	-0.754	-0.073
team[Blackburn]	-0.1583	0.117	-1.350	0.177	-0.388	0.071
team[Blackpool]	0.0036	0.146	0.025	0.980	-0.283	0.290
team[Bolton]	-0.1185	0.115	-1.027	0.305	-0.345	0.108
team[Bournemouth]	-0.0500	0.100	-0.501	0.616	-0.245	0.145
team[Brighton]	-0.4052	0.180	-2.247	0.025	-0.759	-0.052
team[Burnley]	-0.4049	0.113	-3.586	0.000	-0.626	-0.184
team[Cardiff]	-0.4857	0.185	-2.623	0.009	-0.849	-0.123
team[Chelsea]	0.2599	0.069	3.755	0.000	0.124	0.396
team[Crystal Palace]	-0.1907	0.088	-2.174	0.030	-0.363	-0.019
team[Everton]	0.0058	0.073	0.079	0.937	-0.137	0.149

team[Fulham]	-0.1506	0.092	-1.645	0.100	-0.330	0.029
team[Huddersfield]	-0.5956	0.197	-3.023	0.003	-0.982	-0.209
team[Hull]	-0.3581	0.111	-3.232	0.001	-0.575	-0.141
team[Leicester]	0.0578	0.087	0.662	0.508	-0.113	0.229
team[Liverpool]	0.2596	0.069	3.746	0.000	0.124	0.395
team[Man City]	0.4251	0.067	6.325	0.000	0.293	0.557
team[Man United]	0.2416	0.069	3.479	0.001	0.106	0.378
team[Middlesbrough]	-0.6384	0.200	-3.188	0.001	-1.031	-0.246
team[Newcastle]	-0.1404	0.078	-1.797	0.072	-0.293	0.013
team[Norwich]	-0.2856	0.096	-2.967	0.003	-0.474	-0.097
team[QPR]	-0.3248	0.108	-3.000	0.003	-0.537	-0.113
team[Reading]	-0.2074	0.162	-1.279	0.201	-0.525	0.110
team[Southampton]	-0.0684	0.080	-0.855	0.393	-0.225	0.089
team[Stoke]	-0.2654	0.078	-3.406	0.001	-0.418	-0.113
team[Sunderland]	-0.2852	0.081	-3.517	0.000	-0.444	-0.126
team[Swansea]	-0.1847	0.079	-2.332	0.020	-0.340	-0.029
team[Tottenham]	0.2088	0.070	2.985	0.003	0.072	0.346
team[Watford]	-0.2115	0.105	-2.007	0.045	-0.418	-0.005
team[West Brom]	-0.2115	0.077	-2.750	0.006	-0.362	-0.061
team[West Ham]	-0.1014	0.078	-1.307	0.191	-0.253	0.051
team[Wigan]	-0.2389	0.104	-2.297	0.022	-0.443	-0.035
team[Wolves]	-0.2422	0.121	-1.995	0.046	-0.480	-0.004
opponent[T.Aston Villa]	0.3613	0.075	4.806	0.000	0.214	0.509
opponent[T.Birmingham]	0.2793	0.143	1.958	0.050	-0.000	0.559
opponent[T.Blackburn]	0.4629	0.102	4.550	0.000	0.263	0.662
opponent[T.Blackpool]	0.5938	0.126	4.708	0.000	0.347	0.841
opponent[T.Bolton]	0.4351	0.103	4.232	0.000	0.234	0.637
opponent[T.Bournemouth]	0.4052	0.090	4.491	0.000	0.228	0.582
opponent[T.Brighton]	0.2087	0.147	1.421	0.155	-0.079	0.497
opponent[T.Burnley]	0.1150	0.099	1.161	0.246	-0.079	0.309

opponent[T.Cardiff]	0.5376	0.129	4.182	0.000	0.286	0.790
opponent[T.Chelsea]	-0.1151	0.079	-1.449	0.147	-0.271	0.041
opponent[T.Crystal Palace]	0.2116	0.082	2.579	0.010	0.051	0.372
opponent[T.Everton]	0.0782	0.075	1.037	0.300	-0.070	0.226
opponent[T.Fulham]	0.3293	0.085	3.883	0.000	0.163	0.495
opponent[T.Huddersfield]	0.2741	0.142	1.924	0.054	-0.005	0.553
opponent[T.Hull]	0.3467	0.092	3.776	0.000	0.167	0.527
opponent[T.Leicester]	0.2185	0.088	2.493	0.013	0.047	0.390
opponent[T.Liverpool]	0.0499	0.076	0.655	0.513	-0.099	0.199
opponent[T.Man City]	-0.1817	0.081	-2.240	0.025	-0.341	-0.023
opponent[T.Man United]	-0.1710	0.081	-2.122	0.034	-0.329	-0.013
opponent[T.Middlesbrough]	0.1884	0.148	1.273	0.203	-0.102	0.478
opponent[T.Newcastle]	0.3040	0.074	4.128	0.000	0.160	0.448
opponent[T.Norwich]	0.3760	0.083	4.513	0.000	0.213	0.539
opponent[T.QPR]	0.4257	0.090	4.754	0.000	0.250	0.601
opponent[T.Reading]	0.5248	0.129	4.059	0.000	0.271	0.778
opponent[T.Southampton]	0.0938	0.081	1.163	0.245	-0.064	0.252
opponent[T.Stoke]	0.1942	0.073	2.654	0.008	0.051	0.338
opponent[T.Sunderland]	0.2739	0.074	3.699	0.000	0.129	0.419
opponent[T.Swansea]	0.2348	0.075	3.139	0.002	0.088	0.381
opponent[T.Tottenham]	-0.0146	0.077	-0.189	0.850	-0.166	0.137
opponent[T.Watford]	0.3286	0.092	3.563	0.000	0.148	0.509
opponent[T.West Brom]	0.2495	0.072	3.450	0.001	0.108	0.391
opponent[T.West Ham]	0.2900	0.074	3.925	0.000	0.145	0.435
opponent[T.Wigan]	0.4122	0.090	4.573	0.000	0.236	0.589
opponent[T.Wolves]	0.5365	0.099	5.416	0.000	0.342	0.731
home	0.2724	0.022	12.312	0.000	0.229	0.316

Appendix 2

'$\tilde{\alpha}_{Arsenal}$': 1.4423893896288609,

'$\tilde{\alpha}_{Bournemouth}$': 1.1020563638625969,

'$\tilde{\alpha}_{Brighton}$': 0.7528023305988317,

'$\tilde{\alpha}_{Burnley}$': 0.7433504066625719,

'$\tilde{\alpha}_{Cardiff}$': 0.6718472314755046,

'$\tilde{\alpha}_{Chelsea}$': 1.411108814669605,

'$\tilde{\alpha}_{CrystalPalace}$': 0.9564298026309063,

'$\tilde{\alpha}_{Everton}$': 1.1564717740061394,

'$\tilde{\alpha}_{Fulham}$': 1.0016366332716813,

'$\tilde{\alpha}_{Huddersfield}$': 0.5611542389266958,

'$\tilde{\alpha}_{Leicester}$': 1.2101095016484764,

'$\tilde{\alpha}_{Liverpool}$': 1.4115868106526184,

'$\tilde{\alpha}_{ManCity}$': 1.5765701926130324,

'$\tilde{\alpha}_{ManUnited}$': 1.3924094602083803,

'$\tilde{\alpha}_{Newcastle}$': 1.0083483505241735,

'$\tilde{\alpha}_{Southampton}$': 1.083139124841409,

'$\tilde{\alpha}_{Tottenham}$': 1.3589765952394532,

'$\tilde{\alpha}_{Watford}$': 0.9377953529144604,

'$\tilde{\alpha}_{WestHam}$': 1.052792565684589,

'$\tilde{\alpha}_{Wolves}$': 0.9112329995692178,

'$\tilde{\beta}_{Arsenal}$': -1.1506573411988894,

'$\tilde{\beta}_{Bournemouth}$': -0.7466378999787747,

'$\tilde{\beta}_{Brighton}$': -0.9433467852309084,

'$\tilde{\beta}_{Burnley}$': -1.0391436463281694,

'$\tilde{\beta}_{Cardiff}$': -0.6140184400702217,

'$\tilde{\beta}_{Chelsea}$': -1.2664182855494908,

'$\tilde{\beta}_{CrystalPalace}$': -0.9403550405185297,

'$\tilde{\beta}_{Everton}$': -1.0727958997561893,

'$\tilde{\beta}_{Fulham}$': -0.8228914988140964,

'$\tilde{\beta}_{Huddersfield}$': -0.876754040185636,

'$\tilde{\beta}_{Leicester}$': -0.9327228869250347,

'$\tilde{\beta}_{Liverpool}$': -1.0999775222064119,

'$\tilde{\beta}_{ManCity}$': -1.332163819291578,

'$\tilde{\beta}_{ManUnited}$': -1.3222655674334836,

'$\tilde{\beta}_{Newcastle}$': -0.8500185622352655,

'$\tilde{\beta}_{Southampton}$': -1.057762860391228,

'$\tilde{\beta}_{Tottenham}$': -1.1660636548801082,

'$\tilde{\beta}_{Watford}$': -0.8228477166080611,

'$\tilde{\beta}_{WestHam}$': -0.8597553020077582,

'$\tilde{\beta}_{Wolves}$': -0.6168609003546477,

'rho': -0.054032325021451694,

'$\tilde{\gamma}$': 0.2731855529532789

Non-exclusive licence to reproduce thesis and make thesis public

I , Farhad Tahirov

1. herewith grant the University of Tartu a free permit (non-exclusive licence) to:

1.1. reproduce, for the purpose of preservation, including for adding to the DSpace digital archives until the expiry of the term of copyright, and

1.2. make available to the public via the web environment of the University of Tartu, including via the DSpace digital archives, under the Creative Commons licence CC BY NC ND 3.0, which allows, by giving appropriate credit to the author, to reproduce, distribute the work and communicate it to the public, and prohibits the creation of derivative works and any commercial use of the work until the expiry of the term of copyright,

Application of Poisson and Dixon-Coles models on football match outcome prediction and research of a positive return over investment in betting market
supervised by Jüri Lember

2. I am aware of the fact that the author retains the rights specified in p. 1.

3. I certify that granting the non-exclusive licence does not infringe other persons' intellectual property rights or rights arising from the personal data protection legislation.

Farhad Tahirov
17/08/2020