**Predicting results of football games to consistently achieve better accuracy than the one suggested by bookmakers’ odds**

**Research project proposal**

**Michal Borowski**

**Introduction**

My idea for this research project is to analyse information about previous games to predict the results of football games and to consistently achieve better accuracy of prediction comparing with the probability of winning suggested by the bookmakers’ odds. The semi-automatic system would take into account the past results of the football teams with emphasis on the games against the same rivals, with taking into consideration home/away advantage and pre-match performance ratings provided by “whoscored.com”. In order to verify whether the predictions are accurate a simulation is going to be used that will check if this method will consistently “beat the odds” provided by the “365bet” bookmaker company over a long period of time.

**Literature Review with a critical review of key references**

The author of this research proposal is planning to check whether any similar projects have been done in the past. What methodology they used and with what results. Initially the author of this research proposal expected to develop further any projects found, however in practice he found the research that he was not able to fully understand because it was much more complex than the project he planned to do. The author of this research proposal found it demotivating and he is considering to make completely different project which he discussed with the tutor. The “pi-football” project created by Anthony Constantinou et al. (2012) incorporated objective and subjective variables in the model for predicting match outcomes. According to the conclusions of the published document the model which incorporated Bayesian networks was able to consistently “beat the odds” of the various bookmakers, the author also posted the predictions before the games were played which was not practiced with any previous relevant research (for 380 Premier League games around the 2010/2011 season). It also used bookmakers’ odds as a method of verification to see how effective prediction was. While being effective the prediction method used in that research has relatively big limitations when compared to this project proposal which does not require any subjective data input unlike the pi-football which requires the user to input the following parameters based on the experts’ individual judgement:

-team strength

-confidence

-player availability (with regard to their value for the team)

-team spirit

-managerial impact

-head to head bias

-toughness of previous match

-first team players rested during last match

-national team participation

All these factors incorporated into the prediction system make great tool that is merging 2 substantial aspects:

1. Objective statistical data.
2. Subjective information based on individual knowledge and intuition.

From the other side pi-football cannot be automatized and used independently by someone who does not have a way to get required data from an expert. It also requires a lot of effort to provide specific information for each team prior to each game. Additionally the subjective information is specific to an extent that the gathering of it cannot be accomplished for the past games without some extraordinarily detailed database that is already made by either professionals or hard-core enthusiasts (it would be impossible to just remember all the listed subjective parameters for the games that were played few years ago without having “photographic” memory), therefore it would not be possible to easily create long-term testing simulation as described in the 4th point of methodology section of this document. I think it is a significant advantage of the prediction system I am proposing to develop which ideally could be operated by “pressing a button” by anyone and is aimed to work with any major football league regardless of the users’ knowledge about it.

According to Nivard van Wijk (2012) the history of predicting results of football games has its academic origins around early 1980s with a publication of M.J. Maher (1982) which introduced a model designed to predict exact scores by using 4 variables assigned for each team:

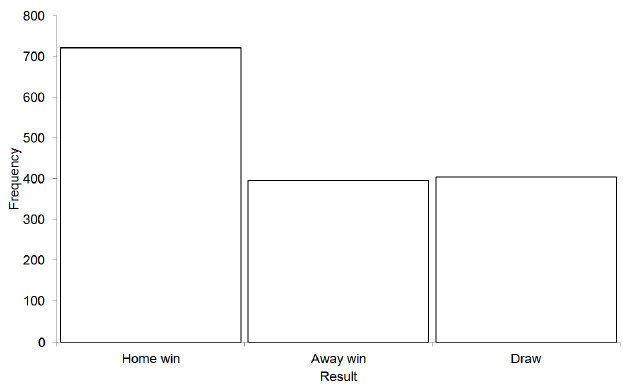
-attack home

-defence home

-attack away

-defence away

Further research carried out by M.J Dixon & S.G Coles (1997) and H. Rue & O. Salvesen (2000) were based on the same model but developed it further by what Nivard Wijk called “join conditional probability law” with inclusion of “correction factor” for some of the scores. The “Soccer Analytics” book presents a graph which shows the distribution of home and away wins over the 1520 Premier League matches from season 2007-2008 until 2010-2011, however from the initial research of the author of this research proposal that distribution changes relatively heavily from year to year.



Graph copied from: "Soccer Analytics: Predicting the outcome of soccer matches" (page 3)

**Methodology (Equipment and test procedures, Research design, and Results Analysis process)**

The author of this research proposal is planning to:

1. Create database by downloading and extracting data from:

-“football-data.co.uk” which has been previously used in similar research such as the one carried out by Razali et.al (2017) and provides information about the games from the past few years in a simple to use CSV format. The information includes:

**•** results

**•** scores

**•** pre-game odds provided by several bookmakers (which will be used to verify how successful is the prediction)

**•** and more (that will not be used)

-“whoscored.com” which provides performance ratings for each team after every game, these 2 rating values (each for every team) will have to be extracted from each link for every Premier League game over the last few years. I’m planning to use an automated program to collect this information.

-“oddschecker.com” – future matches odds

2. Create an algorithm calculating the probability of winning (and providing odds) which will take account of:

-results of the last 19 games (possibly using a higher multiplier for the latest games to increase their significance on the final prediction)

-results of the games between the same teams (called “STR” – same teams ratio within the code/database)

-home advantage multiplier (defined by how often the home teams won/lost/drew over the last few years in Premier League, or possibly it could be defined by the results of specific team when played at home)

3. Compare the odds produced by the algorithm with the odds of the bookmakers (e.g. “bet365”) in order to determine whether it is beneficial to bet. If the odds on some team are lower than the odds provided by “bet365” then the bet would be classified as worth placing.

4. Using the data from the past simulate long-term betting using the betting-system based on the above algorithm. The simulation will “time-travel” few years back and “forget” the latest results, it will calculate the probability of results based on the data before specific events that happened long time ago. Taking advantage that “football-data.co.uk” provides odds from the great past it will be possible to verify the effectiveness of this betting-system as if it was actually used over the last few years. I will create a variable cash amount value which will change with each simulated bet depending on the result.

5. Modify the algorithm and see how it will impact the outcome of the simulation mentioned above. Make sure to use amount of games/bets high enough (thousands) to make sure that the bias of the researcher does not play significant role.

6. Present the results of the simulation on the graph, or think about presenting the simulation itself (possibly with a variable speed). During the presentation explain what is the margin, show an animation of odds added together and divided by 3 and say that the result is the amount of money I would be left with if I bet on all 3 options with £1. Building on that create and present the results of completely random betting simulations (e.g. home teams only, away teams only, draws only, completely random) and compare them with the simulation based on my algorithm.

7. Export all the odds that were bet on during multiple simulations into a separate files and group them into successful and unsuccessful. Use them to create graphs showing the distribution of bets depending on the height of odds comparing successful and unsuccessful bets. Hopefully it should show a clear representation of how effective the predictions were comparing to the predictions of bookmakers.

8. Try to find some literature about this topic which may be difficult to make the algorithm based on the work made by “someone” instead of assumptions of “nobody”. It might be different in this case because 2+2=4 has the same validity when said by “someone” and “nobody” therefore analogically the more complex calculations used in this project may not need to be referenced to any official publisher etc.

9. Find and understand the laws of statistics/probability to provide the value of error and its probability for the long-term testing simulation.

All of the computation and data collection would be done by using a program or set of programs written in ‘python’ except the presentation which would be created in ‘processing 3.3.6’ and PowerPoint.

**Factors that will affect the research (Threats to Validity and Reliability) with control measures**

It is going to be necessary to keep the project as clean and readable as possible. Otherwise it may be not possible to verify its reliability. Every step will have to be documented and the presentation will have to outline the substantial information about the project in a clear and understandable way. Video presentations could be placed in the PowerPoint (if that’s possible) or using other software. A small mistake in the program responsible for the testing simulation could completely corrupt the outcome and invalidate the results therefore each action undertaken by it will have to be examined with scrutiny. The vital point I will be trying to achieve is to portray the used calculations in such way that anyone who tries to recreate them will get the same results, otherwise the whole idea of research will be pointless like a house made of paper. Additionally the correctness of data from ‘football-data.co.uk’, ‘whoscored.com’ and ‘oddschecker.com’ will have an impact on the reliability of the project. Prepare the presentation with a focus on making it clear and understandable from beginning to end instead of trying to “sound clever” and ignoring the fact that it will be presented to the sport-course group.

# References

Anthony Constantinou, N. F. (2012, 02 05). pi-football: A Bayesian network model for forecasting Association Football match outcomes. Retrieved from http://www.eecs.qmul.ac.uk/~norman/papers/pi-football%20%20A%20Bayesian%20network%20model%20for%20forecasting%20Association%20Football%20match%20outcomes.pdf

M.J. Dixon, S. C. (1997). *Modelling association football scores and inefficiencies in the football betting market.* Applied Statistics.

Maher, M. (1982). *Modelling association football scores.* Statistica Neerlandica.

Razali, N. (2017). Predicting Football Matches Results using Bayesian Networks for English Premier League (EPL). *International Research and Innovation Summit (IRIS2017)*. Retrieved from https://www.researchgate.net/publication/319105446\_Predicting\_Football\_Matches\_Results\_using\_Bayesian\_Networks\_for\_English\_Premier\_League\_EPL

Wijk, N. v. (2012). *Soccer Analytics: Predicting the outcome of soccer matches.* Amsterdam: Research Paper Business Analytics. Retrieved from https://beta.vu.nl/nl/Images/werkstuk-wijk\_tcm235-315158.pdf

whoscored.com

https://www.whoscored.com/Regions/252/Tournaments/2/Seasons/6829/Stages/15151/Fixtures/England-Premier-League-2017-2018

https://www.whoscored.com/Regions/252/Tournaments/2/Seasons/6335/Stages/13796/Fixtures/England-Premier-League-2016-2017

https://www.whoscored.com/Regions/252/Tournaments/2/Seasons/5826/Stages/12496/Fixtures/England-Premier-League-2015-2016

https://www.whoscored.com/Regions/252/Tournaments/2/Seasons/4311/Stages/9155/Fixtures/England-Premier-League-2014-2015

https://www.whoscored.com/Regions/252/Tournaments/2/Seasons/3853/Stages/7794/Fixtures/England-Premier-League-2013-2014

https://www.whoscored.com/Regions/252/Tournaments/2/Seasons/3389/Stages/6531/Fixtures/England-Premier-League-2012-2013

https://www.whoscored.com/Regions/252/Tournaments/2/Seasons/2935/Stages/5476/Fixtures/England-Premier-League-2011-2012

https://www.whoscored.com/Regions/252/Tournaments/2/Seasons/2458/Stages/4345/Fixtures/England-Premier-League-2010-2011

football-data.co.uk

http://www.football-data.co.uk/mmz4281/1718/E0.csv

http://www.football-data.co.uk/mmz4281/1617/E0.csv

http://www.football-data.co.uk/mmz4281/1516/E0.csv

http://www.football-data.co.uk/mmz4281/1415/E0.csv

http://www.football-data.co.uk/mmz4281/1314/E0.csv

http://www.football-data.co.uk/mmz4281/1213/E0.csv

http://www.football-data.co.uk/mmz4281/1112/E0.csv

http://www.football-data.co.uk/mmz4281/1011/E0.csv

http://www.football-data.co.uk/mmz4281/0910/E0.csv

http://www.football-data.co.uk/mmz4281/0809/E0.csv

http://www.football-data.co.uk/mmz4281/0708/E0.csv

http://www.football-data.co.uk/mmz4281/0607/E0.csv

http://www.football-data.co.uk/mmz4281/0506/E0.csv

oddschecker.com

http://www.oddschecker.com/football/english/premier-league

http://www.oddschecker.com/football/english/championship

http://www.oddschecker.com/football/spain/la-liga-primera  
http://www.oddschecker.com/football/other/portugal/primeira-liga

http://www.oddschecker.com/football/italy/serie-a

http://www.oddschecker.com/football/other/turkey/super-lig

http://www.oddschecker.com/football/other/netherlands/eredivisie

http://www.oddschecker.com/football/germany/bundesliga

http://www.oddschecker.com/football/france/ligue-1