Czech Technical University in Prague

Faculty of Mechanical Engineering

Department of Instrumentation and Control Engineering

Master Thesis

Student

Bc. Nikita Zhdankin

Supervisor

Ing. Matouš Cejnek, Ph.D.

2020/2021

MASTER'S THESIS ASSIGNMENT

I. Personal and study details

Student's name: **Zhdankin Nikita** Personal ID number: **467461**

Faculty / Institute: **Faculty of Mechanical Engineering**

Department / Institute: **Department of Instrumentation and Control Engineering**

Study program: **Automation and Instrumentation Engineering**

Specialisation: **Instrumentation Engineering**

II. Master's thesis details

Master's thesis title in English:

Modelling of soccer match results for match outcome prediction

Master's thesis title in Czech:

Modelování výsledku fotbalového zápasu za účelem předpovědi závěru

Guidelines:

Thesis objectives:
- research suitable soccer team descriptors
- engineer suitable data-set from available data
- research state-of-the methods of soccer match result prediction
- evaluate and test suitable models for match outcome prediction

Bibliography / sources:

[1] Egidi, Leonardo, Francesco Pauli, and Nicola Torelli. "Combining historical data and bookmakers' odds in modelling football scores." Statistical Modelling 18.5-6 (2018): 436-459.
[2] Inan, Tugbay. "Using poisson model for goal prediction in European football." Journal of Human Sport and Exercise (2020).

Name and workplace of master's thesis supervisor:

Ing. Matouš Cejnek, Ph.D., U12110.3

Name and workplace of second master's thesis supervisor or consultant:

Date of master's thesis assignment: **30.04.2021** Deadline for master's thesis submission: **10.06.2021**

Assignment valid until: _____

Ing. Matouš Cejnek, Ph.D. Supervisor's signature	Head of department's signature	prof. Ing. Michael Valášek, DrSc. Dean's signature

III. Assignment receipt

The student acknowledges that the master's thesis is an individual work. The student must produce his thesis without the assistance of others, with the exception of provided consultations. Within the master's thesis, the author must state the names of consultants and include a list of references.

Date of assignment receipt	Student's signature

Statement

I declare that I have worked out this thesis independently assuming that the results of the thesis can also be used at the discretion of the supervisor of the thesis as its co-author. I also agree with the potential publication of the results of the thesis or of its substantial part, provided I will be listed as the co-author.

Prague ……

Signature ……

Abstract

The goal of the thesis was to evaluate factors affecting football match results, create a rating which would rank the football teams in an objective way and try to predict football match results better than existing methods.

Firstly, several factors were evaluated in a way to what extent each factor affects a football match result. Such factors as home advantage, age, form, motivation, average ball possession and transfer values were evaluated. The results showed that the home advantage is the most important factor in a football match as the team's chances to win almost double if it plays at a home ground.

Then, several predicting models were evaluated. I started with the Elo ranking system, which was improved by implementing the factors which were described previously. This system showed better predictability compared to the Poisson distribution model and betting odds.

Finally, machine learning models were tested using historical data and the Elo ranking system together with several of the factors. The results showed accuracy of 0.535, which means that it would predict 53.5% of the games correctly.

Acknowledgements

Firstly, I would like to thank Ing. Matouš Cejnek, for being my supervisor during my work on this master thesis and helping me throughout the whole year. Also, for the classes of Python, which gave me knowledge during the course.

I want to thank my parents and my sister who were supporting me during the whole period of my studies and especially last year which was stressful due to the pandemic.

I also want to thank all my friends and my girlfriend who were motivating, encouraging, and supporting me.

Contents

List of tables

List of figures

1. Introduction

1.1. Football

Football is a game played with two opposing teams of 11 players and a ball for 90 minutes [1]. The aim of each team is to score the ball into the opponent's goals more times than the opposing team does. If both teams have scored the same number of goals, the game ends in a draw. Any country usually has a domestic league where a fixed number of clubs play against each other. The team with the highest number of points at the end of the season wins the league and 1 or more teams with the lowest number of points are relegated to the lower division and replaced with the best teams of the lower division.

With big money coming into the sport, detailed statistics became crucial in reaching better results. In recent years, new types of data have been collected for many games in various countries, including information on each shot or run made in a match. The collection of this data has made data science one of the most important fields of the football industry with many possible applications: scouting players, improving match tactics, improvements in the training process, injury prevention, betting and many more.

The first part of the thesis is determining which factors and to what extent affect the football game results. The list of the factors would contain several uncontrollable or/and unmeasurable ones, such as luck, mood of the players etc. Michael J. Mauboussin explains the factor of luck in his book "The Success Equation" [2]. He put several kinds of sports on so-called luck-skill continuum where on the left edge it has sports that depend only on luck, such as roulette, and on the right edge there are sports that depend only on skill, such as chess. The graph is presented below:

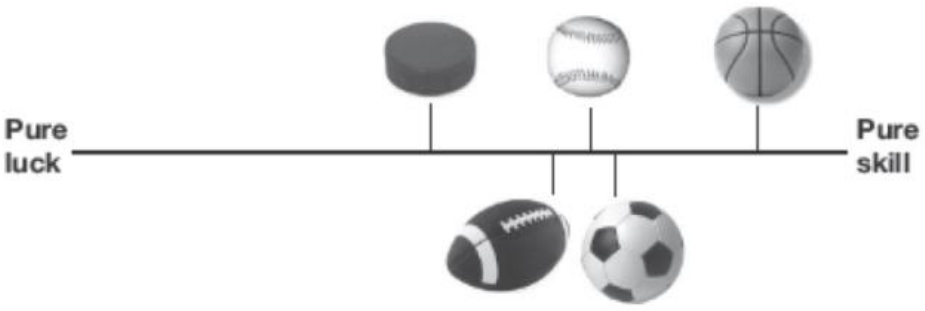

Figure 1 - Luck-skill continuum [2]

As the average number of goals scored in football is between 2.5 and 3, which is relatively small, the sport is strongly affected by the factor of luck. Michael J. Mauboussin explains that the higher the ratio of luck to skill in the sports the larger the sample needed to make reasonable conclusions. For example, an Olympics champion in sprinting would beat an amateur runner literally every time in normal conditions, but when moving left on the scale above, it would require larger and larger sample to understand the contribution of skill. Football is in the middle of the scale.

I will focus on controllable and measurable factors in this thesis. The list of the factors which are going to be evaluated as follows:

- Teams' strength
- Home advantage
- Average ball possession
- Recent form
- Motivation
- Age

The second part of this work is creating a model. It is going to be based on the ELO rating model, but with several adjustments according to the results of the experiments on factors listed above. The Elo rating model will be explained in the chapter 'Elo'.

The third part will be about the Poisson distribution and how well it can predict the football match results. After that, the machine learning models are going to be evaluated in prediction of football matches. Several algorithms are going to be used and my updated Elo ratings are also going to be used.

In the end, I will compare all the created models in order to find the best one. The overall goal of the work is to objectively evaluate football teams as precisely as possible and try to find a model which has the best chances of predicting a football match result correctly.

In my experiments I am going to use databases provided by sources like football-data.co.uk, footystats.com and https://github.com/jalapic/engsoccerdata. The software used to perform the experiments is Python 3.9 and the IDE used is PyCharm by JetBrains with a student license provided by CTU in Prague.

1.2. Odds

Betting is one of the biggest markets in the football industry. Odds represent the probability of an event occurring [3]. For any given event, there are a certain number of outcomes. Take rolling a dice for instance. If someone rolls a dice, there are six possible outcomes. Therefore, if you bet that the person rolls a 'one', there is a 16.67% chance that will happen. The price shown translates into a percentage chance of something happening or not.

The formula for transforming odds to probability is:

$$P = \frac{1}{O} * 100\% \quad (1)$$

Where P is probability and O is odds. The example of calculation is in the table below:

Odds	Probability
1	100%
1.33	75%
1.5	67%
2	50%
2.5	40%
5	20%
8	13%

Table 1 – Meaning of odds

Data analysis is the first and most crucial step in the process of calculating the odds. Bookmakers hire specialists to compile all the data possible and make models to improve predictability of events. They try to get the best tools possible and work with the best software to ensure that they get objective statistical evaluation of each game and the possibilities. These days there is too much information for human beings to follow, that is why bookmakers create mathematical models to set and change odds automatically.

Even though the bookmakers have probably the strongest models for predicting football results, it is not their objective to be correct in the predictions. Their goal is to make as high a profit as possible, therefore, they set the odds in the way they would beat an average better. Bookmakers use advanced algorithms to calculate how much cash flow would be placed on a specific market.

After bookmakers have calculated the odds and the predicted cash flow, they need to post the odds. They adjust the odds with something that is called 'margin'. This factor allows bookmakers to make money. The bookmakers use the margin and provide overall odds that are slightly lower than what they should be. If both outcomes have the same percent probability, then the odds should be even - 2.0. But the actual odds are lower than the conventional ones, which means that they might offer something slightly lower than 2.0 depending on their financial politics instead of even odds. The difference between the odds is the margin itself. The margin varies between different bookmakers – usually from 3 to 10 percent.

In case of an event that could end with 3 different outcomes the margin can be calculated using the following formula:

$$M = \frac{1}{outcome_1} * 100 + \frac{1}{outcome_2} * 100 \frac{1}{outcome_3} * 100 - 100 \quad (2)$$

For example, for the World Cup 2018 Final, the odds were: 2.15 for France' win, 4.23 for Croatia win and 3.08 for the draw, so the margin could be calculated as follows:

$$M = \frac{1}{2.15} * 100 + \frac{1}{3.08} * 100 \frac{1}{4.23} * 100 - 100 = 46.5 + 32.5 + 23.6 - 100 = 2.64\ \%$$

The margin is much lower than usual because the World Cup Final is the biggest market.

The odds can change before the game due to injuries or some other issues within the teams. The odds can also change during the match, because of events, like a red card, player change, injury, penalty, goal, or other events that might change the match's outcome. Another reason why the odds change is because of the cash projections that might affect the bookmakers' profit.

Even though the profit is the bookmakers' goal, as was already mentioned, betting odds can beat any existing mathematical model in predicting football results.

2. ELO

2.1. History

The Elo rating was created by a Hungarian master-level chess player Arpad Elo, born in 1903 he won eight titles of Wisconsin State champion, but he is mostly known by the rating system he created in 1960. Since the United States Chess Federation was founded in 1939 [4], it was willing to apply a rating system which would help members to track individual progress. The Harkness system which was used from 1950 to 1960, used four digits and had a top rating of about 2600, the reason for that was explained by its' creator – Kenneth Harkness – he felt that fewer digits would leave too many players with the same rating. The Harkness system was considered inaccurate; therefore, it was replaced by the Elo rating, and it is still in use by main chess federations including FIDE – World Chess Federation – with minor adjustments. Initially Arpad wanted to decrease the rating of a "strong player" from 2000 to 1000, but it appeared then that it would mean that some weaker player could get a negative rating. That is why Elo has decided to roughly keep the scoring.

2.2. Mathematics

Elo rating's central assumption is that the performance of the player (or team if we are analysing football) is a normally distributed random variable [6], which means that 68% of the values are within +/- one standard deviation of the mean, 95% are within +/- two standard deviations, and 99.7% are within +/- three standard deviations. It is based on the fact that the performance of a given player, or a team is not the same from game to game. It can vary due to multiple uncontrollable factors, but the player's true skill is the mean of his performance random variable.

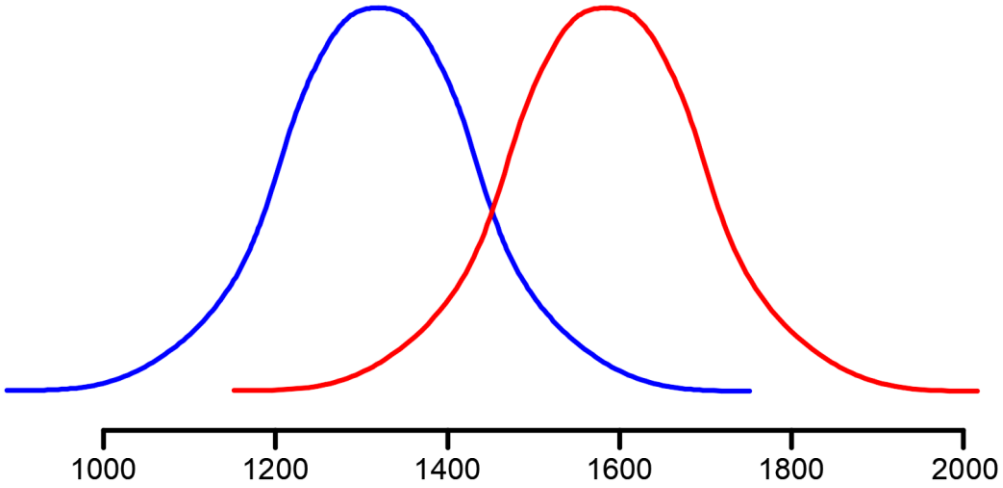

Figure 2 – Elo rating distribution [5]

Graph above shows an example of player 1 (blue line), whose rating is 1300, competing against player 2 (red line), whose rating is 1600. Judging by the graph, player 1 loses the game in most of the cases, but if player 1 overperforms and player 2 underperforms, there is a chance for player 1 to win a game.

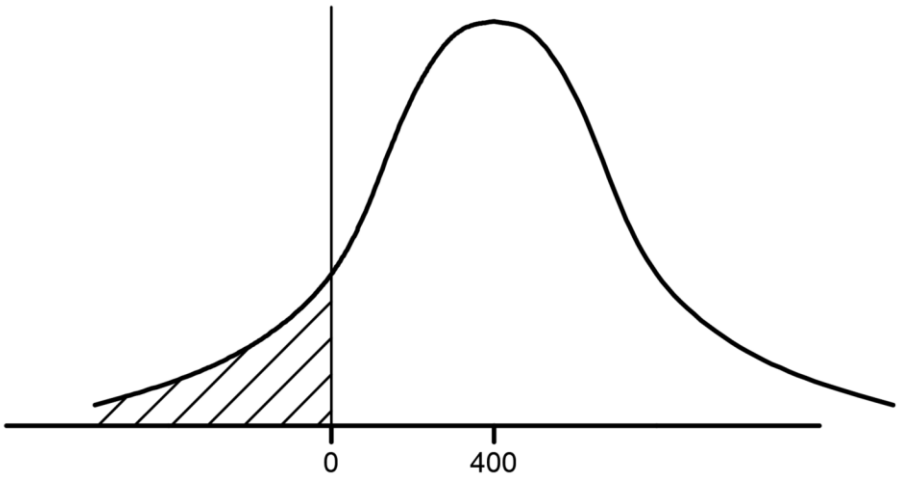

Figure 3 – Elo Logistic curve [5]

The graph above is called a Logistic curve. The Elo rating is designed so that if a player has a rating that is 400 points more than an opponent player, he is 10 times more likely to win a game. That is shown on the graph – the unshaded area under the curve is 10 times greater than the shaded one. Therefore, if a player has a rating that is 800 points more than an opponent, his likelihood to win a game is 100 greater. Putting that into the formula:

$$P_{winA} = 10^{(R_A - R_B)/400} * P_{winB} \quad (3)$$

Where P_{winA} and P_{winB} are the probability of winning of two players, R_A and R_B are their Elo ratings accordingly. This formula can be rewritten as:

$$P_{winA} = 10^{(R_A - R_B)/400} * (1 - P_{winA}) \quad (4)$$

Modifying the formula above it is possible to get the first out of the two formulae that original ELO rating is based on:

$$O_E = \frac{1}{1+10^{\frac{R_B - R_A}{400}}} \quad (5)$$

Where O_E is the expected outcome of the game, which is calculated using ELO rating of teams A and B, and the higher the difference between these two numbers the higher the probability of winning of the team with higher ELO rating. This dependence can be seen on the graph below.

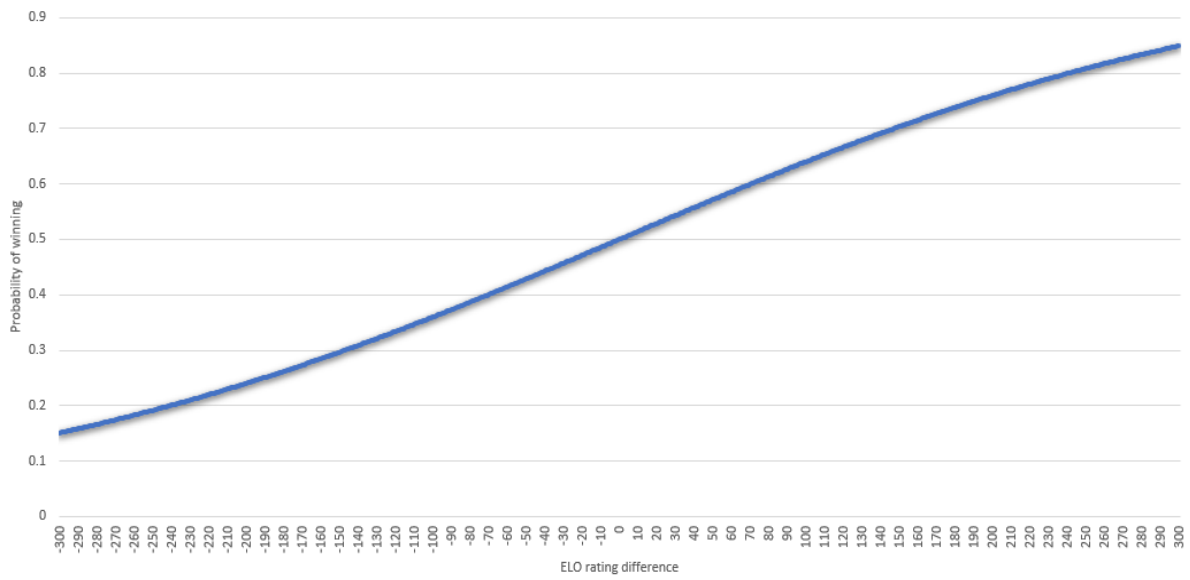

Figure 4 – correlation between Elo difference and win probability

The second formula for original ELO rating is presented below:

$$R_{new} = R_{previous} + K(O - O_E) \quad (6)$$

Where R_{new} is the ELO rating of a team after the match, $R_{previous}$ is the ELO rating of the team before the match, K is weight index, O is the outcome of the match (1 if the team has won the game, 0.5 if the game has drawn, 0 if the team has lost the game) and O_E is the expected

outcome which is calculated by the formula [3]. Using 1, 0.5 and 0 as values for the outcome of the match assume that team A, for example, performed at a better level than team B in the case of scoring more goals, which is not always the case. Therefore, the goal of my model is to take into consideration the cases when a team dominated the game but did not manage to transfer this domination into the goals due to bad luck, biased referee or some other factors.

FIDE, The International Chess Federation, uses the following ranges for the K-factor:

K = 40, for a player new to the rating list until the completion of events with a total of 30 games and for all players until their 18th birthday, as long as their rating remains under 2300.

K = 20, for players with a rating always under 2400.

K = 10, for players with any published rating of at least 2400 and at least 30 games played in previous events. Thereafter it remains permanently at 10.

2.3. Elo model in football

The goal of this work is to modify the formulae so that the updated Elo rating would give much accurate rating of the football teams.

There are several existing Elo models that are applicable for football. http://clubelo.com is the first example, two main formulae for their rating are the same as for the original Elo rating with weight index K = 20. One modification by the website is weighting goal difference. This is implemented via using the following formula:

$$\Delta Elo_{margin} = \Delta Elo_{1goal} * \sqrt{margin} \quad (7)$$

Where:

$$\Delta Elo\_1goal = \frac{\Delta Elo\_1X2}{\sum(\sqrt{margin} * \frac{p_{margin}}{p_{1X2}})} \quad (8)$$

Where ΔElo_1X2 is the number of points that teams exchange after the result of the game is known, p_{margin} is the likelihood for a specific margin, p_{1X2} is the likelihood to win (or lose) by any margin. The sum is over all margins for a win or all margins for a loss.

Another modification for this rating is Home Field Advantage (HFA). ClubElo increases the Elo difference for a match by a certain number of Elo points. Every day and for every country separately, the system compares if home teams won more or less points than away teams. If home teams won more, HFA is increased, if away teams won more points, HFA is decreased according the following equation:

$$HFA \mathrel{+}= \sum \Delta Elo * 0.075 \tag{9}$$

Elo rating for football is also used by the website http://eloratings.net and FIFA World Rankings [7]. Both do not take into account home advantage. FIFA World Rankings do not even have goal margin in the formulae, but both ratings consider importance of the tournament by changing K-factor:

K	Type of the tournament
5	Friendlies played outside the International Match Calendar windows
10	Friendlies played within the International Match Calendar windows
15	Nations League matches (group stage)
25	Nations League matches (play-offs and finals)
25	Confederations' final competitions qualifiers, FIFA World Cup qualifiers
35	Confederations' final competitions matches (before quarter-finals)
40	Confederations' final competitions matches (quarter-finals and later)
50	FIFA World Cup matches (before quarter-finals)
60	FIFA World Cup matches (quarter-finals and later)

Table 2 – K-factor for FIFA World Rankings [7]

There are several problems with the Elo ratings. An increase or decrease in the average rating over all players in the rating system is often called a rating inflation or rating deflation accordingly. This phenomenon has been noticed in chess, where in July 2000 the average rating of the top 100 players was 2644 and by July 2012 this number had increased to 2703.

The rating does not show the strength of the player or team but rather shows the ability compared to other player or team of the championship they compete against each other.

Initially the Elo model has assumed that only a win or a loss is possible. Addition of a draw as a third possible outcome complicates the problem. The simplest way to model a probability of a draw was described in a 1967 article by P. Rao and L. Kupper [8]. The model assumes that the draw happens when the players or teams perform at a similar level. It is supposed that there exists a number D which is the largest difference in strengths displayed

in an individual game that would result in a draw, such that a probability of winning of team A is:

$$P_A = \frac{10^{R_A/400}}{10^{R_A/400}+10^{(R_A+D)/400}} \quad (10)$$

As the chance of winning for team B is calculated by the same formula, it is possible to calculate the probability of the draw by subtracting these two resulting values from 1. It is going to be necessary to calculate the average percentage of games finishing as a draw and to check if there is any correlation between team strength and frequency of draws.

Elo is not the only one rating system created. One of the most famous rating systems is the so-called Glicko system created by Mark Glickman [9], who modified Elo rating by introducing the rating deviation variable RD. The goal was to avoid cases when some players played less frequently than others, therefore decreasing precision of the prediction by the model. As my model predicts football matches, all clubs play the same amount of games, so introducing rating deviation would not make any difference. The Glicko system also applies a rating volatility σ to each player which defines how player performance varies from his original rating.

J.Lasek, Z. Szlavik and S. Bhulai in their paper "The predictive power of ranking systems in association football" [10] compared current FIFA Rankings for national teams with several types of Elo ratings, least squares rating etc. They found that any of the Elo ratings would give more accurate results than the official FIFA Rankings. The authors also point out that it should be noticed by FIFA as the rankings affect the football competitions, for example, the draws for international tournaments take into consideration national teams' rating. Therefore, it might be manipulated, for example, playing friendly matches with high probability of increasing one team's rating.

2.4. 3 points for a win

As it has been mentioned in a previous chapter, the Elo model uses the following points system: win is 1 point, draw is 0.5 points and loss is 0 points. But in reality, in professional football 3 points are awarded for a win, 1 point is awarded for a draw and the same 0 points for a loss.

Originally, two points for win were introduced when the football leagues were starting, and it seemed like a reasonable decision [11]. That went on for more than 90 years, but in the 1980s, football was facing serious problems with attendance. The main reason for that seemed to be that the teams value the draw point too much and would not risk going for a win, making football seem boring. Crowds had almost halved in the early 1950s, and it was clear that something had to be done. In the year of 1981 one of the English broadcasters proposed the reward for win to three points. He received some criticism, that it would make a winning team, after scoring a goal, want to sit on their lead even more than previously, but the system of 3 points for a win was introduced anyway.

Research showed that applying three points for a win to every season going back to the second world war, and in each case the champions would remain the same. In the five seasons before the switch, there were an average of 133.0 draws per season in what was then the league; in the five seasons after, there was an average of 113.4. The change seems also to have promoted more attacking play. In the five seasons leading up to it, home teams averaged 1.60 goals per game and away teams 1.01; afterwards home teams averaged 1.64 and away teams 1.07. That is a small change, but optimists could even argue that away teams had become proportionally more attacking – suggesting they were less prepared to play for draws.

A 2005 study by the economists Luis Garicano and Ignacio Palacios-Huerta deeply analysed the effects the points system change had on football [12]. In this study the researchers explained that it is not only about a decrease in the number of draws. The number of matches decided by a large number of goals declined. Measures of offensive effort such as shot attempts on goal and corner kicks increased while indicators of sabotage activity such as fouls and unsporting behaviour punished with yellow cards also increased. Attacking effort increased approximately the same degree as the number of fouls increased resulting in an unchanged amount of goals scored. One more reason for that is assumed to be that the losing team is less willing to score, because the one point awarded for a draw is worth less in a new system.

In conclusion, two points for a win is going to be used in my thesis in order to comply with the Elo model, which assumes that a win is worth twice as much as a draw.

3. Statistics

Two big statistics analysis measurements that are going to be used a lot in this work are correlation and regression. In this part of my thesis I am going to explain them in detail.

Correlation is used for a quick and simple summary of the direction and strength of the relationship between two or more numeric variables [13]. Regression is used for prediction, optimizing, or explaining a number response between the variables, how one variable influences another one.

Both measurements can quantify direction and strength of the relationship, but only regression is able to show cause and effect, predict and optimize the relationship.

Correlation can be spotted when a change to one variable is then followed by a change in another variable, whether it be direct or indirect. Variables are considered uncorrelated when a change in one does not affect the other. If two variables are moving in opposite directions, like when an increase in one variable results in a decrease in another, this is known as a negative correlation. Knowing how two variables are correlated allows for predicting trends in the future. The main purpose of correlation, from the point of view of correlation analysis, is to find out the association or the absence of a relationship between two variables. The degree of association is measured by a correlation coefficient, denoted by r. It is sometimes called Pearson's correlation coefficient after its originator and is a measure of linear association. If a curved line is needed to express the relationship, other and more complicated measures of the correlation must be used.

Unlike correlation which can be defined as the relationship between two variables, regression shows how they affect each other [14]. Regression analysis helps to determine the functional relationship between two variables so that it is possible to estimate the unknown variable to make future projections on events and goals. Regression formula is presented below:

$$y = a + b * x \tag{11}$$

Where a refers to the y-intercept, the value of y when x = 0 and b refers to the slope. In case of several variables affecting one, multiple regression analysis is used.

The formula for multiple regression is:

$$y = a + b * x_1 + c * x_2 + \cdots + n * x_n \quad (12)$$

The main difference between two measurements is that regression defines how x causes y to change, and the results are going to change if x and y are swapped. In case of correlation, x and y are variables that can be interchanged, and results stay the same.

4. XG

Expected goals (xG) is a predictive model used to assess every goal-scoring chance [15], and the likelihood of scoring a goal. Each shot on goal made is evaluated between 0 and 1. It means that if a shot is 0.43 xG, out of 100 attempts the average player would score approximately 43 goals.

There is no commonly accepted way to calculate xG of the shot. But it is known that most of the sports analytics laboratories use following factors to evaluate a goal-scoring chance:

- Distance to the goal
- Angle to the goal
- Which part of the body was a shot made with
- Type of the shot (open-play, free-kick, corner, or counterattack)
- Number of defenders between an attacker and a goal
- Assist type (long ball or through ball)

Analytics companies use big historical data in order to construct a model with high accuracy.

The main assumption in the model is that every player has approximately the same finishing skill in theory. In practice there are some players that would constantly score more goals than xG model predicts them to score. But even the best finishers would not score more than 20% higher than their xG, therefore, it would not affect my model.

xG is a sufficient way to indicate whether results are based on sustainable factors like a constant creation of good chances, or whether it is down to aspects such as luck or good

goalkeeping. For example, if a team is scoring more than it is predicted by the xG model, it is expected that this team's results would get worse due to regression to the mean. Surely, the difference between goals and xG might just tell about the style of the team, but the difference would still decrease over time. This factor I have evaluated in the following experiment.

The data for the last 6 years for top-6 championships (Italy, England, Spain, France, Germany, Russia) was gathered from the understat.com – 36 seasons in total. The goal of the experiment was to find out which parameter predicts the final position of the team better after half of the season: goals, xG or combination of both.

In order to use xG in table prediction it is necessary to introduce another metric called xPoints (XP). It is basically just the expected number of points teams are going to get with the xG number they got in a game between each other. For example, team A plays against team B and according to xG, the possible outcomes of the game are team A win is 50%, draw is 30%, team B win is 20%. XP is calculated [16] in this case as follows:

$$XP = [3^* \times (\text{Possible Outcome of a win})] + [1 \times (\text{Possible Outcome of a Draw})] \quad (13)$$

$$XP_A = (3^* \times 0.5) + (1 \times 0.3) = 1.8 \quad (14)$$

$$XP_B = (3^* \times 0.2) + (1 \times 0.3) = 0.9 \quad (15)$$

* in my experiment I use 2 points for a win system

The win probabilities are calculated using a Monte-Carlo Simulation.

Monte-Carlo Method is a mathematical technique [17] used to estimate the possible outcomes of an uncertain event. It was invented by John von Neumann and Stanislaw and was named after a famous casino city due to the element of luck that is a core to the modelling approach, similar to the game of roulette. Monte-Carlo Simulation works in the way that it predicts a set of outcomes based on an estimated range of values versus a set of fixed input values. It builds a model of possible results by using a probability distribution, such as a uniform or normal distribution, for any variable that has inherent uncertainty. It recalculates the results multiple times, each time using a different set of random numbers between the minimum and maximum values. Usually this exercise can be repeated thousands of times to produce a large number of likely outcomes.

In order to find out whether or not the xG model can help predict the football game result The procedure of the experiment consisted of getting data from the first half of each of 36 seasons analysed and predicting the final position of each club based on one of three factors: points, xPoints and 1:1 combination of both. Once again, 2 points were awarded for the win and 1 point was awarded for the draw. The process of the experiment has shown below on the example of Bundesliga 2014 season:

Club	Final number of Points (FP) predicted using mid-table Points (P)	FP predicted using mid-table xPoints (XP)	FP predicted using combination of P and XP (PXP)	Actual FP	Difference between predicted FP by P and actual FP	Difference between predicted FP by XP and actual FP	Difference between predicted FP by PXP and actual FP
Augsburg	36	31.8	33.9	34	2	2.2	0.1
Bayer Leverkusen	42	42	42	44	2	2	2
Bayern Munich	62	54.2	58.1	54	8	0.2	4.1
Borussia Dortmund	22	35.8	28.9	33	11	2.8	4.1
Borussia M.Gladbach	40	38.8	39.4	47	7	8.2	7.6
Eintracht Frankfurt	34	33.4	33.7	32	2	1.4	1.7
FC Cologne	28	27.8	27.9	31	3	3.2	3.1
Freiburg	26	28	27	27	1	1	0
Hamburger SV	26	25.6	25.8	26	0	0.4	0.2
Hannover 96	34	28.4	31.2	28	6	0.4	3.2
Hertha Berlin	26	24	25	26	0	2	1
Hoffenheim	38	31.4	34.7	32	6	0.6	2.7
Mainz 05	30	29.8	29.9	31	1	1.2	1.1
Paderborn	30	29	29.5	24	6	5	5.5
Schalke 04	38	30	34	35	3	5	1
VfB Stuttgart	26	24.6	25.3	27	1	2.4	1.7
Werder Bremen	26	28.4	27.2	32	6	3.6	4.8
Wolfsburg	48	42.4	45.2	49	1	6.6	3.8
				Average	**3.67**	**2.68**	**2.65**

Table 3 – Bundesliga, 2014, final number of points predictions

After finding an average of the results of the experiments on all 36 seasons using the script ‘xpvspoints.py’ the following results were obtained:

Difference between predicted FP by P and actual FP	Difference between predicted FP by XP and actual FP	Difference between predicted FP by PXP and actual FP
4.878	5.825	4.873

Table 4 - average final number of points predictions

As it was observed, xPoints showed the worst prediction of the future results while combination of using both points and xPoints showed better prediction, but the prediction based only on points is statistically insignificantly worse, even when different proportions in the combinations are used. One more problem with xG is that the sample is relatively small, therefore, it is decided not to use xG in the improved Elo model.

One more problem with the xG model is that open data usually provides just the final value of xG for a single match, for example, team A finished the game with several shots worth 0.7

xG in total and team B with 1.8 xG. But there is a correlation between how dangerous each of the shots were. In order to show this correlation, I have conducted an experiment based on the Monte-Carlo Simulation.

In my experiment I have assumed that team A has done 4 shots: 0.3 xG, 0.2 xG, 0.1 xG and xG. These numbers were fixed during the simulations. Team B shots were an independent variable, the sum of the xG was the same – 1.8 xG, only the average xG per shot and number of shots were changed. The number of simulations was 100000 for each experiment. The results are presented below:

Average xG per shot	Percentage of matches won
0.9	78.05
0.6	71.68
0.36	67.82
0.13	65.16
0.075	64.53

Table 5 – average xG per shot

The results of the experiment prove that the sum of the xG for the match does not show the full picture of the match and such stat as xG per shot would increase the accuracy of the model.

5. Factors affecting a result of football match

5.1. Home advantage

One of the most important affecting the football match and its result is so-called “home advantage”. It means that if a team plays “at home”: in the city the club is based on, on the stadium the team plays half of its game during the season. In this part of the thesis I am going to discuss what exactly gives the home advantage and to what extent it affects the result of the football match.

For the experiment data was used from 2014 to 2018 years. Data for 2019 and 2020 was excluded as it was hugely affected by the crisis caused by coronavirus. As stadiums stopped

allowing supporters to come to football matches, it has illuminated one of the main factors causing home advantage – fans support.

It is necessary to find the percentage of games won by home teams. This was done by the code named 'home_win_percentage.py'. It calculated the combined win rate from all championships involved in the experiment, excluding last two years data for the reasons described above. The results for the home teams were as following:

Average win rate is 49.62 %

Average draw rare is 25.72 %

Average loss rate is 24.66 %

As seen above, it can be surely said that home advantage is something very significant in any football match. Home teams are winning approximately twice more often than the away teams. In football leagues and cups this effect is illuminated as usually in leagues when each team must play with each team, they play 2 times: home and away.

In cups with a play-off system, the final is usually one match which is played on a 'neutral' pitch in order to avoid the case when one team has home advantage and another one doesn't.

It is also a known fact that 6 out of 18 FIFA World Cups were won by the hosting countries [18]. It might be affected by the fact that earlier the World Cups were hosted by the most powerful teams. But it is impossible to ignore such performances like Sweden (17th in the 1958 ELO Rating) getting to the final, Chile (23rd in the 1962 ELO Rating) getting to the semi-final, Mexico (36th in the 1970 ELO Rating) getting to the quarter-final, USA (53rd in the 1994 ELO Rating) getting to the round of 16, South Korea (23rd in the 2002 ELO Rating) getting to the semi-final, Russia (44th in the 2018 ELO Rating) getting to the quarter-final.

Reasons are considered to be different each time, for example, in South Korea in 2002 referees have made some questionable decisions leading Korea to the top-4 of the tournament. [19] Mexico is several thousand meters above the water level, therefore lower atmospheric pressure, which affects players performance. [20] The other case is the overperforming of the Russian national team in 2018 which has probably been caused by the motivation given from their fans' support. [21]

The studies on evolutionary psychology compare home advantage with ‘the protective response to an invasion of one’s perceived territory’, it is proven by the testosterone levels: for those who play at home it is significantly higher than The studies also show that the home advantage is stronger at the time when a league has just been created and decreases over time. [22] This fact may be ignored for this experiment as all considered leagues have decades of history.

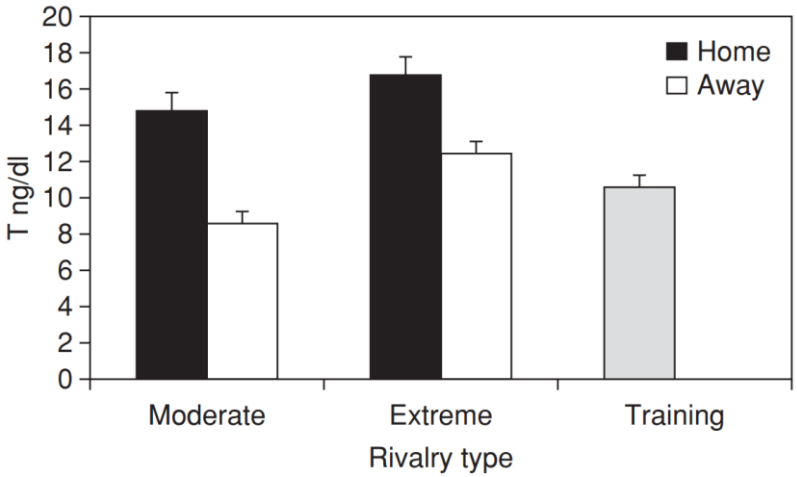

Figure 5 – testosterone levels of football players [22]

Even though the home advantage is a combination of many factors, it is still possible to evaluate some of them.

The experiment is divided to several parts to investigate which factor has more impact on the home advantage, these factors are:

- Distance which an away team has travelled to get to the football pitch
- Match attendance
- Geographical factors, such as area of the country and climate.

The data were taken from the most popular and most rich football nations, which are: Argentina, Austria, Belgium, Brazil, China, Denmark, England, Finland, Germany, Greece, Ireland, Italy, Japan, Mexico, Netherlands, Norway, Poland, Portugal, Romania, Russia, Scotland, Spain, Sweden, Switzerland, Turkey, USA

Taking into consideration less popular championships might affect the results of the experiment. For example, Nigerian Football League, season 2012/2013 [23]. The conditions

are so unique, that top 16 teams in the table have not lost a single home game in the whole season. The reasons for that are the following: referees are under threat if they make decisions against the home team, even if those decisions are correct; violent home crowds; dangerous and exhausting travel etc.

Data for the experiment was taken from the following websites: https://www.football-data.co.uk/data.php and https://github.com/jalapic/engsoccerdata

The first main factor analysed is travel distance. It is assumed that the more time a visiting travel spends on travelling the harder the game will be for them. When the travel distance is high, it might also be the case of different climates or different time zones, which makes it even more difficult to play for the away team.

The goal of the experiment was to find correlation between average travel distance for the club throughout the season and the club's average performance at home field.

The first step in the experiment was to get the data for chosen championships. The first part of the data is the travel distance of each club. The procedure was as following:

- Creating .csv documents with the list of the clubs that were playing at least one season in the championship during the time between 2014 and 2018. The code is in the file 'clubs.py'
- Creating .csv documents with the list of the clubs and the cities their home stadiums are in. The data were taken from https://en.wikipedia.org
- Creating .csv documents with the list of the clubs, their home cities and longitude and latitude, the geographical data was received from https://latitudelongitude.org
- Creating .csv documents with the list of clubs, their home cities, longitude and latitude, and the distance (in kilometres) the team travels throughout one season. It is calculated as the sum of distances between one and other cities in the file 'calculating_km.py'
- To compile average travel distance and average team performance at home file 'calculate_win_percentage_club.py' is used
- To visualise the data for each championship the file 'travel_vs_win_clubs.py' is used.

The findings of the experiment show that there is a correlation between how much the team travels to play away or how much the opponent team travels to play this team and how well the team performs at home during the season.

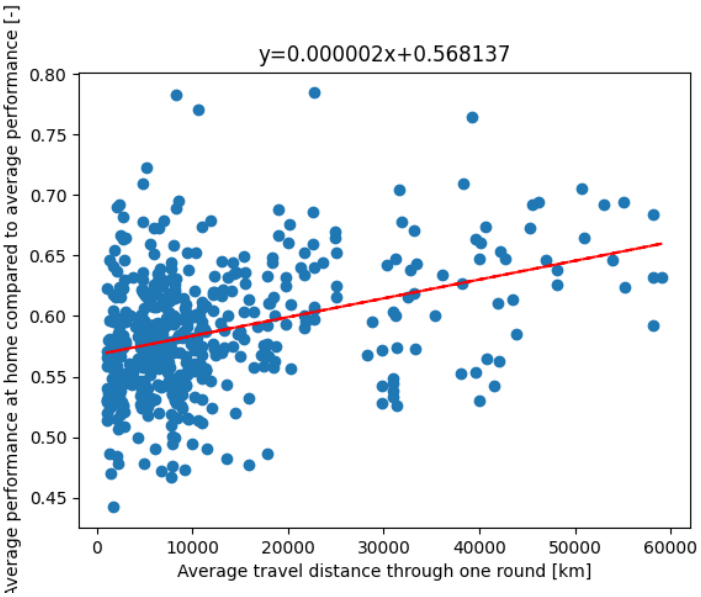

Figure 6 – travel distance affecting home performance

The x-axis of the graph is average travel distance throughout the season, the y-axis is average performance at home compared to average performance. The graph shows that there is a correlation, which means that the more the team travels the better its' performance at home rather than away.

Some more significant graphs are presented below.

The first example is UD Las Palmas, a football club from Spain. It is based in the Canary Islands on the Atlantic Ocean, significantly far from mainland Spain, therefore having significantly higher at home performance comparing to other clubs in the league.

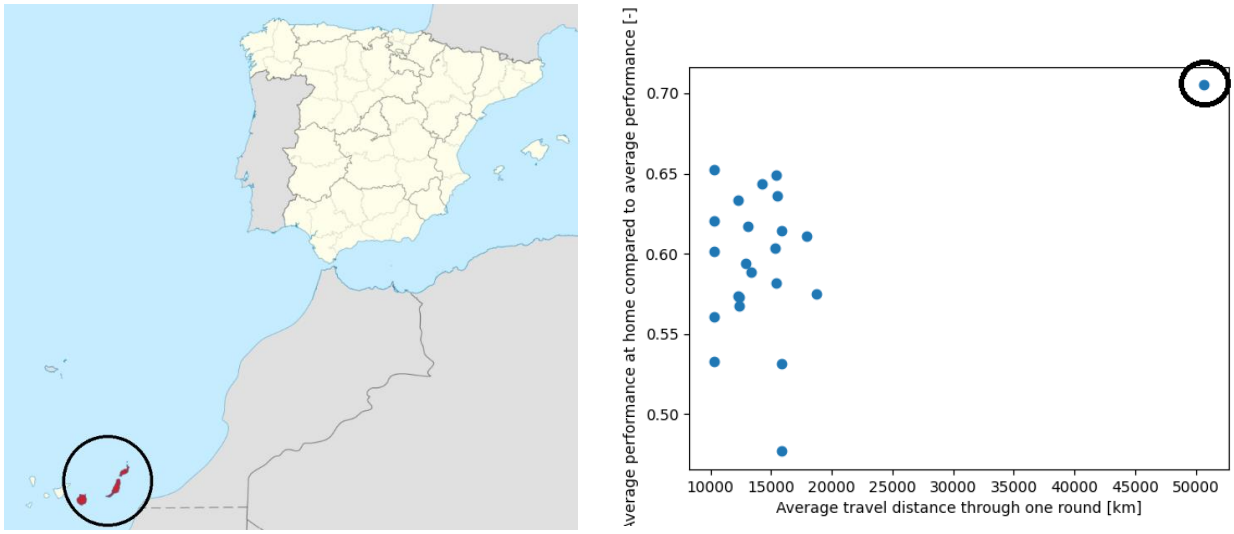

Figure 7 – UD Las Palmas home performance and location on the map [24]

The first picture above shows where the UD Las Palmas is based on the map, the second graph shows where this club is on the graph of average travel distance versus home performance.

The second example are two clubs from Portugal: CS Maritimo and CD Nacional. Both clubs are based on the island of Madeira located approximately 600 kilometres away from mainland Portugal. Therefore, these two clubs have significantly higher home performance due to long travel comparing to other clubs in the league.

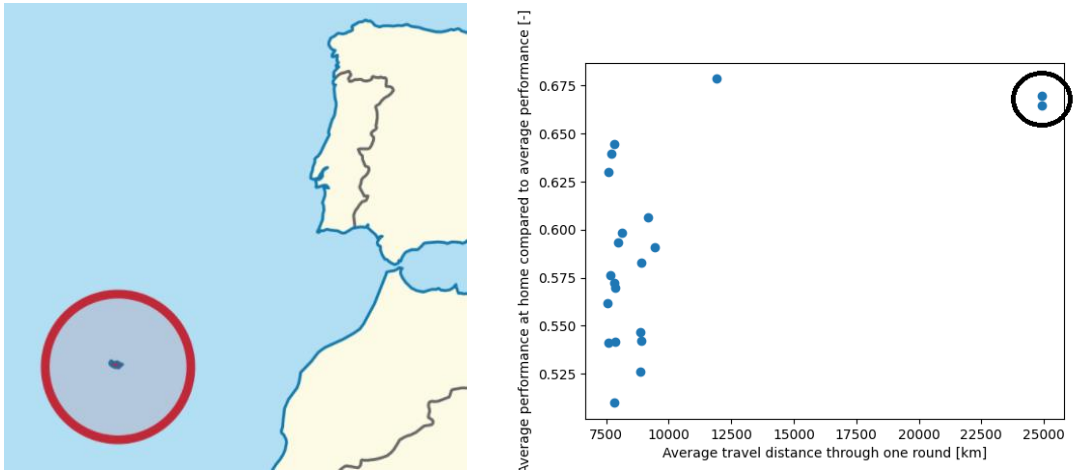

Figure 8 – CS Maritimo and CD Nacional home performance and location on map [25]

The first graph above shows where CS Maritimo and CD Nacional are based on the map, the second graph shows where these clubs are on the graph of average travel distance versus home performance.

The next factor evaluated is the average attendance throughout the season.

The procedure of this experiment was as following:

- Creating .csv documents with the list of the clubs that were playing at least one season in the championship during the time between 2014 and 2018. The code is in the file 'clubs.py'
- Creating .csv documents with the list of clubs with the average attendance of the home stadium. The data were taken from the website https://www.transfermarkt.co.uk
- Creating .csv document with average attendance for each championship using 'avg_attendance_for_each_championship.py' file
- The data is visualised then by using the file 'attendance_vs_win_clubs.py'

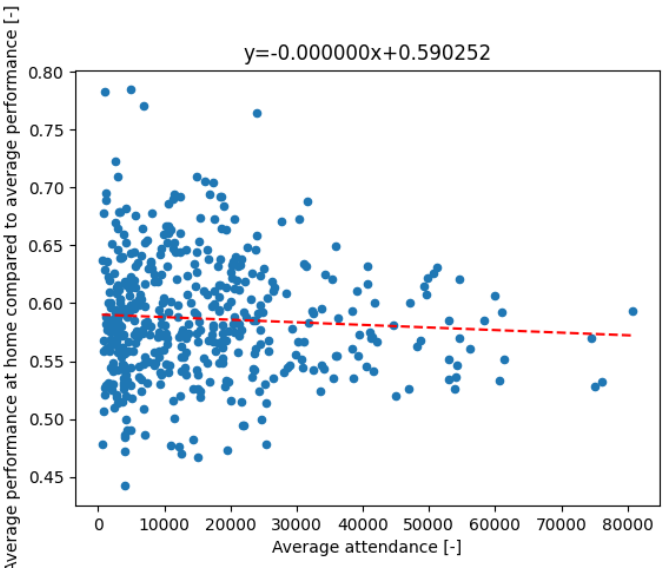

Figure 9 – Average attendance affecting home performance

Unfortunately, the results of the experiment did not show any correlation between average attendance and performance at home field. It even showed slight negative correlation, which is not what was expected before the start of the experiment. The explanation may be that higher attendance means a more expensive stadium which in turn means a richer

football club and therefore a stronger team. And stronger team would have a good away performance as it would be easy for them to beat any opponent even at away field.

After experiment failure, it was decided to verify that at least teams with widely known great support would show some correlation between attendance and home performance. Several teams with 'loudest' [26] fans were listed, with average home performance and average home performance for their championship.

Place	Club	Country	Average relation of home points to away points in championship	Club relation of home points to away points
1	Galatasaray	Turkey	1.504	1.423
2	Besiktas	Turkey	1.504	1.25
3	Fenerbahce	Turkey	1.504	1.248
4	Red Star	Serbia	No data	
5	Borussia Dortmund	Germany	1.567	1.456
6	PAOK	Greece	1.767	1.306
7	Celtic	Scotland	1.138	1.085
8	Boca Juniors	Argentina	1.398	1.252
9	Feyenoord	Netherlands	1.54	1.301
10	Liverpool	England	1.606	1.229
11	Club America	Mexico	1.481	1.23
12	Rangers	Scotland	No data	
13	Wisla	Poland	1.404	1.387
14	Milan/Inter	Italy	1.506	1.326 / 1.231
15	Penarol	Uruguay	No data	

Table 6 – home performance of the most supported clubs

Surprisingly, even the football teams with the most active fan base usually perform worse at home according to the graph above. Not even a single club could perform at home better than on average. The sample is very small and objective to make a conclusion that an active fan base affects negatively, but it is worth considering.

The correlation that could not confirm the initial assumption that higher attendance would cause higher home advantage, could be explained in several ways.

In later years as the financial side of football has become more crucial and clubs have started to gain as much income as possible, one of the solutions was to bring more people to the stadiums. Therefore, stadiums became more family-oriented with new kinds of entertainment during matchday, consultations with the community of supporters, flexible pricing and creating family only areas in the stadium.

For the biggest and most famous football clubs like Barcelona or Arsenal, a huge part of people coming to matches are now tourists who came to visit Spain and the UK respectively, for whom visiting the football match is like an attraction. The categories described above would not make a difference for home or away teams, as it is considered that they would not support any of the teams playing.

There were also some other parameters checked if they affect the home advantage. The following parameters were picked:

- Area of the country, which should correlate with the average distance travelled for the clubs.
- Average attendance
- Average temperature of the air throughout the year

The code for visualising the data is in the file 'other_correlations.py'

The results did not show any significant dependency of the home advantage on any of the factors above. There is a slight correlation for each of the factors. But it is hard to take it into consideration due to the small size of the correlation. But it proves the point that the home advantage is a huge combination of different factors which is also different for each country and even for each game.

5.2. Age

Another factor that might affect the result of a football match is the age of the players. With the modern technologies currently, it is possible for an athlete to give the best performance for much longer time than previously, due to multiple researches in medicine, dietetics, and others. In football this change is obvious, in the Champions League, main European football tournament, an average age has increased from 24.9 to 26.5 years between 1992/1993 and 2017/2018 seasons.

Barça Innovation Hub in their article 'The Influence of Age on Footballers' Performance' analyse several studies in order to understand more about how the players' age affect one's performance level [27]. They concluded that there is a clear loss of physical performance for players over 30 years old comparing to younger players. This was explained by the fact that

older players' total distance covered is 2% less on average than younger players. Much higher difference was found for the number of sprints and fast races - decrease by 21% and 12% respectively.

However, the technical and tactical performance was found to be better in older players. The percentage of successful passes is 3% to 5% higher in players over 30 compared to younger players. It is possible that the decrease in physical performance of younger players is compensated by an improvement in other skills such as decision making and game intelligence.

The article concludes that the combination of youth and maturity in a squad of players may be the best formula for success of a team. It would seem necessary to individualise as much as possible the players' preparation according to their age: they don't all need the same training to reach the best version of themselves. Besides, it is crucial to adapt players' positions and roles in the game to bring up the best qualities from each player on the field.

In order to confirm the conclusions stated above, the following experiment was conducted. The data of average age of a starting XI were taken from https://www.transfermarkt.co.uk, and was plotted against the number of points the team gets. Average age of a team was adjusted to an average age of the league and the number of points was calculated as an average number of points per game throughout the season. The results of the experiment showed no correlation between age and results of a team, correlation coefficient was calculated as -0.00176. The script used was 'ppgvsage.py'.

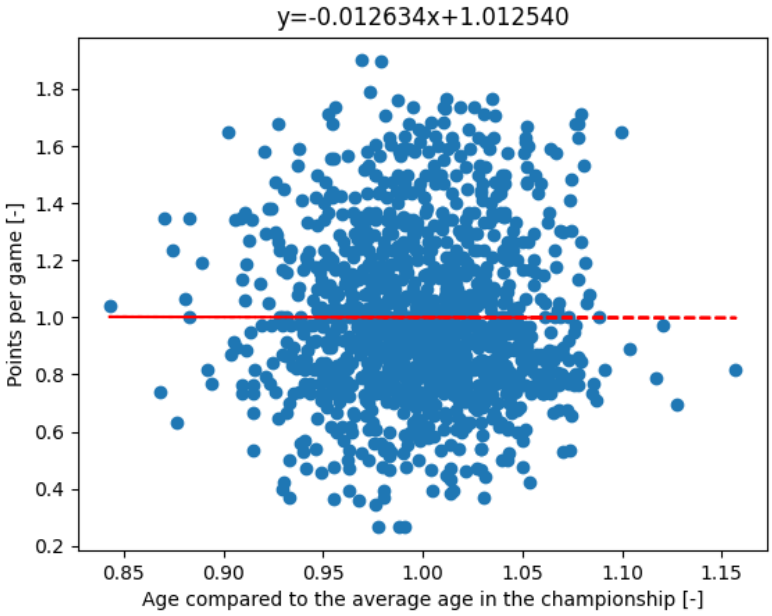

Figure 10 – average age affecting average performance

It was also decided to divide the leagues into three categories according to the UEFA Countries Coefficients. The experiment was then conducted for each of the three groups. The results showed -0.0487 correlation coefficient for the first category (the best leagues), -0.00505 for the second category and 0.0535 for the third category. The scripts are named 'tierone.py', 'tiertwo.py' and 'tierthree.py'. It may be assumed that the worse the quality of the league, the more success a team can get by having older and experienced players in the squad. But the correlation is not sufficient to include the age into the final model.

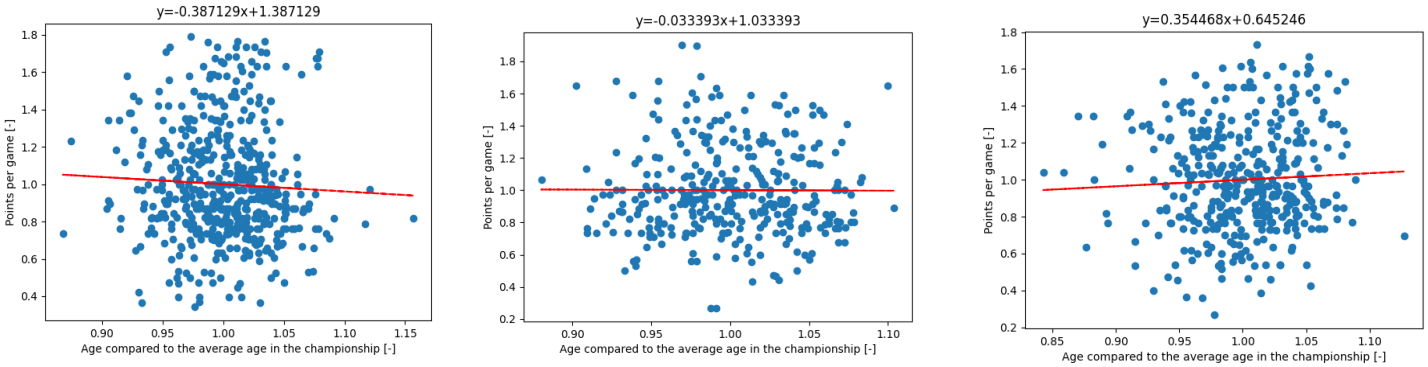

Figure 11 – average age affecting performance of teams in different championships

5.3. Form

One of the factors that might affect the result of a football match is the so-called “form” of the teams. The concept is based on the idea that if a team is playing well or above expectation, it is going to continue playing well due to increased self-confidence [28]. There are not a lot of studies on that subject, but the book “Myths and Facts About Football” suggests that this concept cannot be proved by numbers. In particular, the book study was based on players, not teams, which is more important for creating a prediction model. Therefore, the experiment on teams’ form is going to be conducted.

The study on players’ form might be questionable as it was discussed previously, football is more affected by the factor of luck due to the low amount of shots on goal during the game. Therefore, the study on basketball might be more significant, as the concept of form is based on the factors that are also applicable for basketball. The number of attempts to score a goal in basketball is high compared to other team sports, therefore, the data should be more significant. For example, in the season 2018/2019 (the last season which was not affected by COVID-19) the highest number of attempts for a player in English Premier League was 137 by Mohamed Salah, and in the NBA the number was 1911 by James Harden.

The concept of form in basketball is called “Hot Hand” [29]. It was studied by Gilovich, Vallone and Tversky in 1985. It was based on the assumption made by Kahneman and Tversky in 1972 that people tend to overestimate representativeness of a small sample. Scientists assumed that the Hot Hand is a misconception that would not be proved by the numbers. After their assumption was not disproved by a research, they offered 26 people to throw 100 balls in the net each from the positions with average success rate of approximately 50% and then measured the number of successful and unsuccessful attempts after a streak of successful shots. As the results showed that the number of shots scored after successful streaks is about 50% the study concluded that Hot Hand was a myth.

After Big Data has entered basketball, the Hot Hand concept was revisited by Miller and Sanjurjo in 2018 who wanted to verify if critics of 1985 study were right [30]. Besides several obvious facts like small samples, poor experimental conditions, lack of defensive players etc, scientists noticed that the mathematical approach itself was wrong. Gilovich, Vallone and

Tversky did not take into consideration a concept called streak selection bias which could be explained by a coin flip. If a coin is flipped three times, there are eight different outcomes:

Three-flip sequence	Proportion of heads outcome on recorded flips
TTT	-
TTH	-
TH**T**	0
H**TT**	0
TH**H**	1
H**T**H	0
H**HT**	0.5
H**HH**	1

Table 7 – streak selection bias [30]

According to the table above, the chance of getting head after head is 5/12, not ½ as it could be expected. This bias could be prevented by larger samples. After Miller and Sanjurjo re-evaluated the results of the 1985 experiment, they concluded that the hot hand is not a myth. The explanation was as follows: Gilovich, Vallone and Tversky calculated the average scoring of those 26 participants, which appeared to be 47%. Hot hand (scoring after 3 consecutive successful shots) had 49% probability and "Cold hand" (scoring after 3 consecutive unsuccessful shots) had 45% probability. But the error was to calculate the average probabilities when participants had different numbers of streaks. For example, if one player made 10 shots and scored 4 of them (40%) and another made 5 and scored 3 (60%), their average would be (7/15 = 45%), not ((40+60)/2=50%). The right method would be to calculate an average of streaks, which would give the hot hand advantage up to 20%.

The conclusion above should be now proved experimentally for football. The experiment's procedure was as follows:

Data for several championships for multiple seasons was structured in a way that for each football match there are several parameters for both teams: average points scored per game before the current match, average points scored per game for the last 4/6 matches. These numbers are chosen for the following reasons: the number should be even to avoid home advantage, 2 matches sample is too small, 8 matches sample is too big and might be affected by games played over a month ago. The number in the table represents the relation between average points scored if a team is in good form and average points scored

throughout the season. It is also necessary to notice that the points were calculated as 2 points per win for the reasons explained in the 2 points per win part of the thesis.

			Good form win rate over season win rate (4 matches)						Good form win rate over season win rate (6 matches)					
			1.2+		1.4+		1.6+		1.2+		1.4+		1.6+	
Country	First year	Last year	Relation	Sample size	Relation	Sample size	Relation	Sample size	Relation	Sample size	Relation	Sample size	Relation	Sample size
Argentina	2012	2019	1.058	113	1.146	78	1.253	28	1.057	99	1.122	43	1.107	13
Austria	2012	2019	1.019	64	1.054	40	1.007	23	0.963	46	1.074	25	1.064	9
Belgium	2000	2019	0.985	248	1.056	144	1.114	77	0.993	195	1.048	88	1.183	37
Brazil	2012	2019	0.956	120	0.990	75	1.117	39	0.951	106	1.053	53	0.924	15
China	2014	2019	1.082	58	1.207	28	1.139	18	1.022	45	1.097	21	1.307	9
Denmark	2012	2019	0.967	72	1.033	47	1.185	24	1.001	58	0.976	32	1.248	8
England	2000	2019	0.964	1602	0.999	1143	1.069	591	0.970	1399	1.015	736	1.075	275
Finland	2012	2019	0.954	69	1.000	35	1.002	18	0.982	58	0.952	25	0.970	8
Germany	2000	2019	0.966	542	1.057	335	1.160	163	1.001	437	1.070	206	1.107	76
Ireland	2012	2019	0.998	63	1.073	41	1.051	23	1.036	44	1.176	24	1.333	10
Italy	2000	2019	0.977	307	1.011	193	1.099	96	0.957	264	1.072	122	1.211	51
Japan	2012	2019	0.978	106	1.085	58	1.236	32	1.019	82	1.061	41	1.151	13
Mexico	2012	2019	1.032	106	1.070	65	1.129	29	0.997	91	1.127	43	1.116	15
Netherlands	2000	2019	0.962	272	1.000	163	1.021	91	0.983	226	1.094	105	1.185	46
Norway	2012	2019	1.057	78	1.032	47	1.066	25	1.057	63	1.045	34	0.899	12
Poland	2012	2019	0.981	90	1.014	59	1.165	29	1.034	75	1.182	40	1.264	15
Portugal	2000	2019	0.953	245	0.959	146	1.007	69	0.964	199	0.994	83	1.209	27
Romania	2012	2019	0.992	88	1.096	54	1.144	31	1.008	70	1.091	40	1.155	21
Russia	2012	2019	0.927	83	1.054	45	1.119	24	0.992	66	0.935	29	0.876	13
Scotland	2000	2019	0.962	673	1.022	434	1.100	224	0.969	571	1.023	270	1.080	116
Spain	2000	2019	0.977	327	1.004	215	1.104	112	0.970	274	1.079	126	1.119	46
Sweden	2012	2019	1.008	78	0.988	44	1.072	24	1.080	63	1.097	27	1.209	9
Switzerland	2012	2019	0.986	55	0.986	39	1.124	24	0.991	48	0.961	25	1.425	8
USA	2012	2019	0.917	117	0.977	75	0.928	33	0.915	105	1.021	38	1.117	9
		Average	0.983	5576	1.033	3603	1.094	1847	0.994	4684	1.054	2276	1.139	861

Table 8 – results of the experiment on form affecting team's performance

The results of the experiment show that if a team's recent points per game are 1.4 times higher than one throughout the season, it is 3-5% more likely to win a football match. Accordingly, if a team's recent points per game are 1.6 times higher than one throughout the season, it is 9-14% more likely to win. In conclusion, form appears to be a significant factor in a football match, therefore, it will be implemented in my prediction model.

5.4. Motivation

Motivation is another factor that affects the result of a football match. It refers to a psychological feature that encourages a person to stay active and interested in a specific goal, which in the case of a football game is to defeat the opposing team by delivering the best performance possible [31]. The process that leads to this state of mind is complex and briefly shown on the picture below:

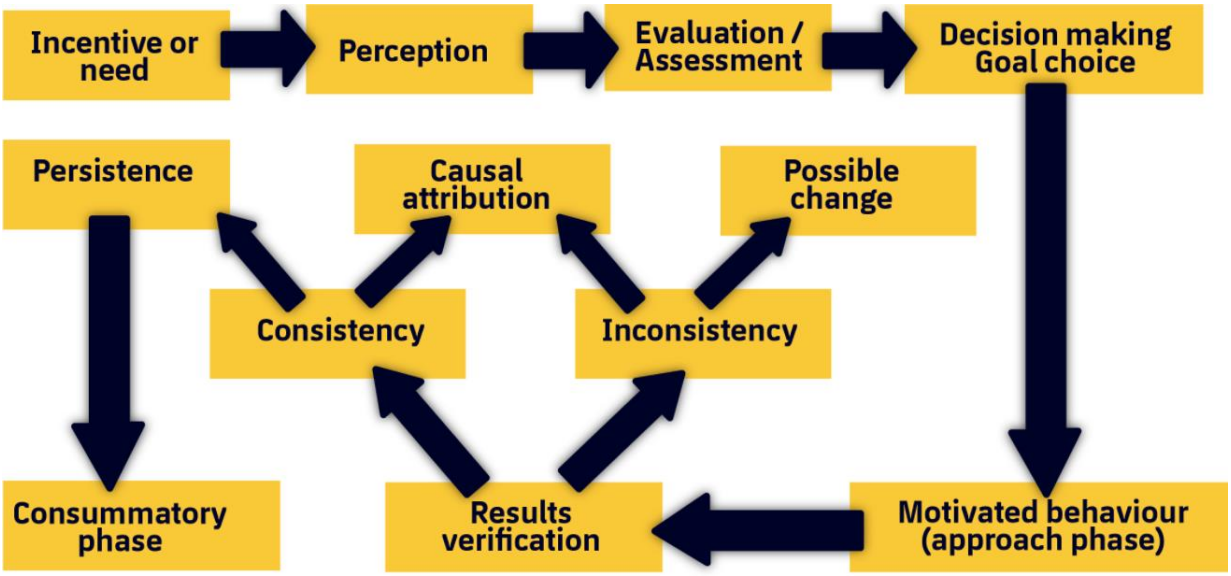

Figure 12 - how motivation affects performance [31]

It is assumed that motivation comes from two types of impulses: internal and external. Internal are the ones that come from the footballer's inner self and are usually related to self-realisation. External are related to factors such as the recognition of others or the achievement of goals during a competition. Therefore, there are several factors that can discourage an athlete, such as low self-esteem caused by, for example, the comparison to other players, or their own doubts regarding the competition. All this can completely destabilise sports performance. According to a 1978 study by Harter, when a person is good at a sport, this situation promotes self-efficacy and, ultimately, motivation to be better. However, if the attempts to improve fail, the motivation will be increasingly lower, and it may even make the player quit the sports practice. As the internal motivation is difficult to transform into numbers and, therefore, impossible to measure, in this experiment I will focus on external motivation, which is the competition. For example, when playing in the league with several other teams, a player would look at the league table and see that any result of the game would not affect the team's position in the league table, one's external motivation would not be as strong.

Assuming that a team would perform worse when there is nothing that can change in the league table whatever the result of the match will be, the experiment is conducted to check whether teams actually perform worse, when they are safe from changing places (4 or more points away from the teams above and below in the league table).

The script for the experiment is 'motivation.py' and it finds cases when the team is safe from any change in the table, it then compares an average number of points per game before the game and the actual points gained from the game which is supposed to be insignificant.

The results show that usually a team would perform 3.91% worse in case of insignificant game compared to usual with a sample of 8682 football matches. It is more interesting that if a game happens in the second half of a season, the difference increases to 5.27% with a sample of 6740 matches. Accordingly, if a game happens in the fourth quarter of a season, the difference becomes 8.4% with a sample of 2872 games.

In conclusion for this chapter, the numbers received from the experiment can be helpful in the model to prevent overestimating the teams that have their place in the table secured.

5.5. Ball possession

The next factor that affects a result of a football match is assumed to be ball possession. That is the percentage of time that a team has a ball in possession. Therefore, if a team A plays against a team B:

$$BP_A + BP_B = 100\% \tag{16}$$

It is a common conception that if a team has higher ball possession, it is more likely to win a game. This concept was mostly popularized by Johan Cruyff, Barcelona manager in the early 1990s [32]. He had the idea that if his team kept possession and created movement, working the ball around the pitch, they would generate more chances and consequently this would lead to success. It is possible to argue with that point of view, as there is probably a wrong causal relationship. It is not the possession that brings good results, but top teams that have a good win rate, prefer to keep the ball, not allowing opponent teams to have a lot of chances to score a goal. The good example is English Premier League champions of 2015/2016 Leicester City, which finished a season with a lowest 44.8% on average. So my assumption that average ball possession is not showing the team's ability to win games but more the team's style.

Soccerment analytics team has explored how possession leads to winning the games in their article “Ball Possession in European Football” [33].

According to their conclusion, the number of goals scored by a team is directly correlated with the number of shots on target. Number of shots in its turn is correlated with the number of accurate passes. Finally, the number of accurate passes depends on the length of passes. The shorter the average length of passes – the higher the number of accurate passes.

Overall graph is shown below.

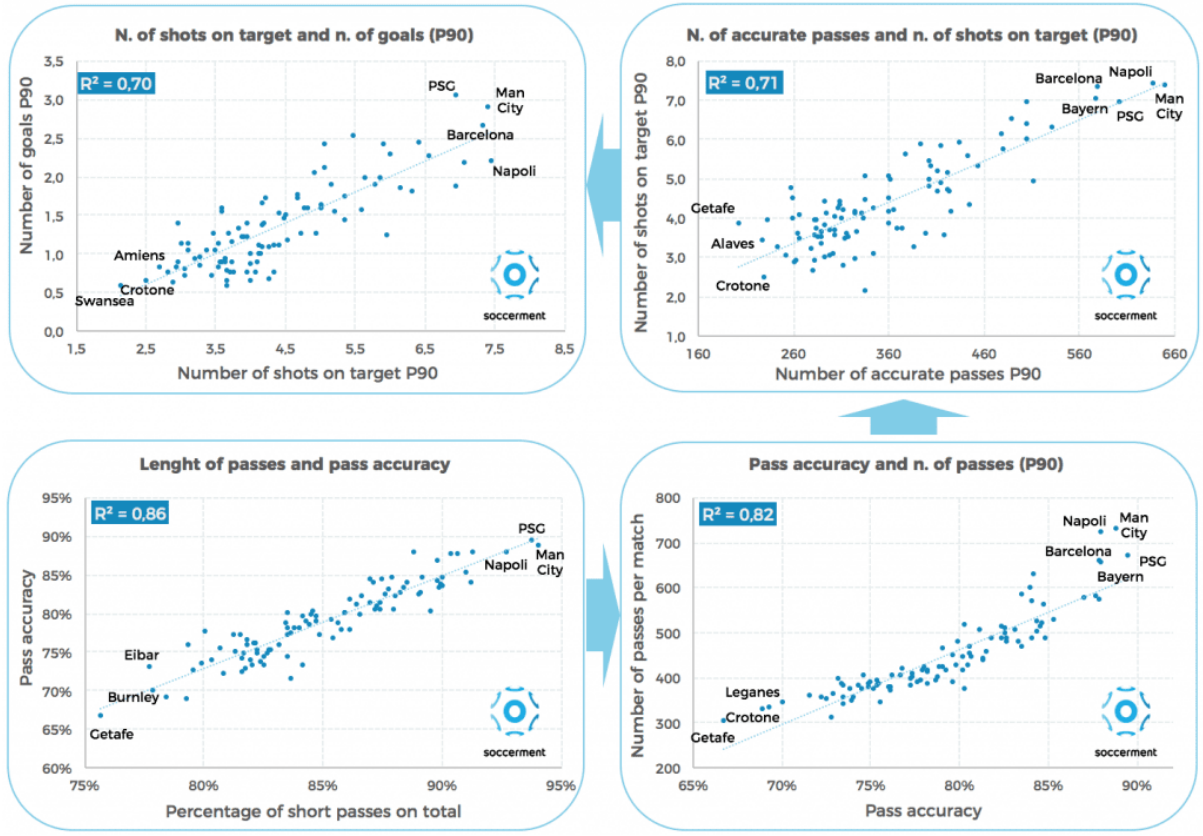

Figure 13 – ball possession causing correlation with the number of goals scored

In this chapter the goal of the experiment is to check several correlations: firstly, I want to compare how team with average {>44,44-47,47-50,50-53,53-57,57<} % of ball possession plays against teams with average {>44,44-47,47-50,50-53,53-57,57<} % of ball possession comparing to average performance. For example, I might find that some team A that has the ball 38% of the time on average, has an average win rate of 20%, but when playing against teams with ~50% average ball possession, this team might have a higher win rate. The

reason for that may be that team A is used to play without the ball, therefore, practicing deep defence and counterattacks. So, when their opponents are forced to play with the ball (so that their ball possession is, for example, 10% higher than usual) they are not familiar with that style of play and they might make more mistakes. This hypothesis will be checked in the following experiment.

Data used in the experiment were taken from the website https://footystats.org/. The championships and seasons are presented in the table below

Championship	First season	Last season
Belgium	2015/2016	2018/2019
England	2011/2012	2018/2019
France	2011/2012	2018/2019
Germany	2011/2012	2018/2019
Greece	2014/2015	2018/2019
Italy	2011/2012	2018/2019
Netherlands	2014/2015	2018/2019
Portugal	2014/2015	2018/2019
Russia	2013/2014	2018/2019
Spain	2011/2012	2018/2019
Turkey	2014/2015	2018/2019

Table 9 – data used for the experiment

Only the top-tier leagues are presented in the experiment. Seasons 2019/2020 and 2020/2021 are excluded due to effects from the Covid-19 pandemic. The total sample consists of 22178 matches.

The results of the experiment are presented in the table below

	0-44	**44-47**	**47-50**	**50-53**	**53-56**	**56-100**
0-44	1.0331	1.1859	1.1117	0.9580	0.7066	0.4497
44-47	1.2613	1.2319	1.1224	0.9270	0.7119	0.4399
47-50	1.2566	1.1956	1.1024	0.9324	0.7623	0.5191
50-53	1.1588	1.2108	1.1102	0.9340	0.6979	0.5449
53-56	1.1896	1.1364	1.0620	0.9916	0.7841	0.6248
56-100	1.1340	1.1255	1.0270	0.9587	0.8484	0.6627

Table 10 – ball possession experiment results

Where the numbers in the table mean

$$b = \frac{points\_gained\_in\_the\_game}{average\_points\_gained\_during\_the\_season} \quad (17)$$

The table shows that for any teams the performance is better when they play against teams with lower average ball possession. So, the results did not give any useful insights. It shows that the teams with lower average possession consistently play worse than teams with higher average possession.

It is also necessary to check if bookmakers tend to overrate teams with higher possession.

	0-44	**44-47**	**47-50**	**50-53**	**53-56**	**56-100**
0-44	-	-2.7267	-4.8115	-19.563	-23.776	-0.2426
44-47	-0.3677	-	-1.2754	-12.166	10.0071	-0.3796
47-50	1.08536	-1.469	-	-1.2478	0.82707	0.0727
50-53	-1.991	-1.0145	-1.9816	-	-1.2627	-0.5238
53-56	0.29453	0.68545	-3.2106	-5.6122	-	-3.3465
56-100	-0.7645	-3.7157	-18.2	-8.5808	-0.9494	-

Table 11 - ball possession experiment results

The numbers represent the profit if betting on the win of the team on Y axis against a team on X axis. Unfortunately, this graph also barely gives any insights. Therefore, the average possession factor is not going to be included in my model.

5.6. Red cards

Red cards can be shown to any of the players on the field during a game if one violates the rules of the game. The red card results in sending off the field for the rest of the game and forbidding the player to play in the next game. This leaves the team in a vulnerable position as they are left with 10 men on the pitch, giving the opposition side the advantage. A player can receive a red card in several cases, for example, if he is guilty of serious foul play or violent conduct, if uses offensive or insulting language or gestures, if he denies the opposing team a goal or an obvious goalscoring opportunity by handling the ball, or he receives a second yellow card in the same match, which is awarded for some less serious violations[34].

The main point is that a red card shown in the game might affect the result of the match in an unexpected way. In order to prove it the following experiment was conducted. The sample of 34879 games was divided by three categories:

- Games when a home team got at least one red card
- Games when an away team got at least one red card
- Games when no red cards were shown to neither of the teams

Games when both teams got at least one red card both were eliminated due to low sample size. After that, each of the categories was divided into three groups:

- Games when a home team was a favourite
- Games when an away team was a favourite
- Games when there were no favourites

Therefore, forming 9 different groups, for each of them a percentage of home win / draw / away win were calculated. It is important to notice that a team being a favourite was determined by betting odds. As my model is not yet available to determine the favourite by a rating, therefore, it was decided to use predictions which supposedly would beat any other mathematical prediction model – betting odds. So, a team was named a favourite when an odd for a win was less than 2.0. The results of the experiment are presented on the table below:

Home team gets a red card					
Home team is a favorite		No favorite in the game		Away team is a favorite	
Home team wins	301	Home team wins	274	Home team wins	32
Draw	246	Draw	370	Draw	79
Away team wins	286	Away team wins	691	Away team wins	287

Away team gets a red card					
Home team is a favorite		No favorite in the game		Away team is a favorite	
Home team wins	1154	Home team wins	1063	Home team wins	114
Draw	269	Draw	498	Draw	99
Away team wins	121	Away team wins	264	Away team wins	117

No red card in the game					
Home team is a favorite		No favorite in the game		Away team is a favorite	
Home team wins	7367	Home team wins	4814	Home team wins	552
Draw	2527	Draw	3841	Draw	718
Away team wins	1601	Away team wins	4205	Away team wins	2149

Table 12 – red card experiment results

The results show that when a team is a favourite and its opponent gets a red card the win percentage is 74.2% while the win percentage of a favourite in a usual game is just 63.8%. This outcome shows that if a team's opponent gets a red card, the chances of winning for the team increase drastically, and this fact should be implemented into my model.

5.7. Transfer values

In order for a player to play for a club in an official tournament it is required that this player has an acting contract signed with the club [35]. The contract usually states the conditions of partnership between a player and a club. But most importantly, it forbids a player to play for another teams. In order for team A to have some player from team B to play for them, it is required to pay a so-called transfer fee. It is also possible to just wait till a contract expires and then sign a new contract with a player without paying a transfer fee, but it was not always like that.

In 1990, Jean-Marc Bosman, Belgian footballer, was coming to the end of his contract at Belgian football club RFC Liege and got an offer to play for French club Dunkirk [36]. But at that times a player could not leave at the end of their contract unless that club agreed to let him go on a free, or that club received an agreed fee from a buying club. Liege has asked for a million dollars in fees from Dunkirk, which was much higher than his market value. The dead fell through and Bosman's salary was cut by 70% and he started a several years of legal battle against a football association. In 1995, the so-called Bosman ruling finally passed and allowed players to leave the clubs after the end of the contracts, therefore, increasing the market exponentially. For example, the transfer fee record was 16.5 million euros before the Bosman ruling, and currently the record is at 222 million euros, showing an increase of more than 10 times.

Transfers in football play an important role. The table below is a clear example of this.

	Season 2019/2020		
	Team	"Domestic" players	Transferred players
1	Borussia Dortmund	0	11
2	Paris Saint-Germain	1	10
3	Bayern Munich	3	8
4	Manchester City	0	11
5	Juventus	0	11
6	Liverpool	2	9
7	Barcelona	2	9
8	RB Leipzig	0	11
9	Valencia	3	8
10	Real Madrid	1	10
11	Tottenham Hotspur	2	9
12	Atalanta	0	11
13	Atlético Madrid	2	9
14	Napoli	1	10
15	Lyon	2	9
16	Chelsea	3	8
	Sum	22	154

	Average number of "domestic" players	Average number of transferred players
	1.375	9.625
In percentage	12.5%	87.5%

Table 13 – percentage of domestic players in top football clubs

The table analyses teams that played in the play-offs of the major international competition – UEFA Champions League. The number of players who graduated from the clubs' youth academies is much lower than those who came from other clubs for a transfer fee.

Simon Kuper and Stefan Szymanski in their book "Soccernomics" are explaining that the correlation between how much a club spends on wages to its players and the club's results is much higher than how much money the club spends on transfers.

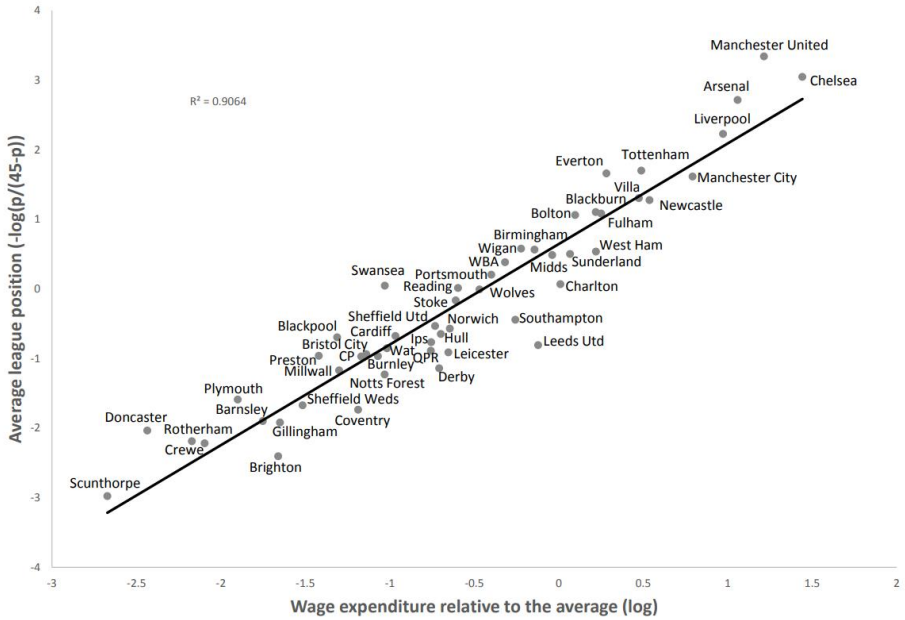

Figure 14 – how players wages affect team's performance [38]

Authors later say that they don't believe that if some club took a random bunch of players, and doubled their salaries, they would suddenly play twice as well. It's not that high pay causes good performance. Rather, high pay attracts good performers. The correlation is obviously weaker on a distance of one season and especially on a distance of one football match. But the conclusion is that the more a club pays its players in wages, the higher it will finish, but what a club pays for them in transfer fees doesn't seem to make much difference.

The main problem with using wages in the model is that there is no open source that would publish that information as almost all football clubs do not announce what salary each player gets. Therefore, I have to choose to use transfer values which are issued by the website https://www.transfermarkt.co.uk which is known for its objectivity and reliability in computation of transfer values. They are calculated using opinions of dozens of independent experts. These values are used in different ways, for example, in August 2020, Antonio Sese, formerly a consultant at Valencia, Spain, went to a city court to appeal against FC Valencia boss Peter Lim and player agent Jorge Mendes, who guided countless players from Portugal. Both are accused of various charges. Among others, Mendes and Lim are accused of using player transfers for their own personal gains. The state prosecutor, using Transfermarkt market values, has now rebuffed those charges [39].

As was told in the interview by one of the Transfermarkt experts [40], there is no specific formula for calculating transfer values, but there are a lot of factors affecting it, such as statistics, demand, age, scout reports etc. Manual decisions are considered to be more effective than neural networks in Transfermarkt.

The transfer values itself might not help to predict the football results but it can be used for teams entering the league. Each year adjacent divisions in each country are exchanging one or more teams. Teams that finish the season in the bottom of the table are relegated and teams that are in the top of the table are promoted to the next division. As my model is going to be used only for the highest leagues of several championships due to lack of data, it would require some data from the new teams that are being promoted. That is where the transfer values are going to be useful, as points are not going to be as significant, because the teams prefer to update most of the squad during an off-season after they are promoted.

6. Improved Elo model

6.1. Formulae

The formulae for Elo rating were described in the Elo chapter of the thesis. In order to improve this model, formulae were adjusted to be more accurate and predict the outcome better than the original model.

The formula for the rating change after the game is:

$$\Delta = (R - E) * k \tag{18}$$

where:

- R is actual result as a value between 0 and 1
- k = 20 * (n/4 - i) for i < n/4 or:
- k = 20 * (n/2 - i) for i < n/2 for new teams.

where n – number of games per season, i is the current gameweek.

The k-factor is chosen to be higher at the beginning of the season as it is the most unpredictable time of the season due to player transfers, manager replacements and possible changes in the play styles of different teams. Therefore, the rating is going to be changed more drastically.

The formula for expected outcome of the match is:

$$E = \frac{1}{1+ 10^{\frac{-dr-HA-M-RC-F}{400}}} \tag{19}$$

where:

- dr is rating difference
- HA is home advantage (88)
- M is motivation (30 for last quarter of the season, 15 if earlier)
- RC (100)
- F is form (30 for great (1.6+) form, 15 for good (1.4+) form)

6.2. Model

The process of getting the results for my model for each championship is:

1. Getting the Elo ratings from the clubelo.com website by using the script in `getelotocsv.py` file. The script accesses the clubelo's API for a date of 1st June each year – when all championships are finished – and saves the data into a .csv file accessible for further manipulation.
2. Calculating the expected results of the games according to the clubelo.com model using the script in `elo.py` file. The script calculates the expected outcome of each game of the season according to the formula provided by clubelo.com and the rating in the beginning of the season. The expected outcomes are also compared with the actual outcomes to spot the effectiveness of the model.
3. Calculating the expected results of the games according to the betting odds using the script in `odds.py` file. The script calculates the expected outcome of each game of the season according to the odds provided by football-data.co.uk. As was mentioned in the Odds chapter of the thesis, bookmakers include the margin in their odds, therefore, the odds were adjusted in the way it is a number between 0 and 1 in this way the same as the Elo model. The expected outcomes are then compared with the actual outcomes to define the effectiveness of the odds.
4. Getting the transfer values from Transfermarkt.co.uk using the script in the `transfervalues.py` file. The script scrapes the data from the Transfermarkt.co.uk and saves the data into a .csv file so that it would be possible to access the data later.
5. Due to minor differences in football clubs' names in Transfermarkt.co.uk and football-data.co.uk it is necessary to change the naming so that it would be the same in all files in order to avoid any data misinterpretation. This process is completed using the script in the `getclubnames.py` file.
6. Calculating the expected results of the games according to my model. The script is chosen between these three files: `my_newratingsfromelo.py`, `my_newratingsfromprevious.py` and `my_newratingsfromtv.py`. The choice depends on the data available. In case of the first season available for the analysis, the initial ratings are chosen from the clubelo model as it is the base for my model. In case of Transfermarkt.co.uk data being available, the ratings are kept the same as

it was the last season and the ratings of the newly promoted teams are going to be calculated according to the regression model provided by the sklearn python library. In other cases, the ratings of the promoted teams is going to be equal to the lowest rating of the teams in the league.

7. Also, the final ratings for my model are saved into the .csv file after each season using the script in the `my_finalratings.py` file.
8. Finally, the three models – clubelo, odds and mine – are compared using the script in the `comparison.py` file.

The experiment was conducted for several championships for multiple seasons. The results showed that my model has better prediction results than Elo model and bookmaker odds given that the margin is 0. The results of all games were adjusted to the goal difference and were placed on the scale between 0 and 1 in order to comply with the Elo formula.

Year	My model result	Elo model result	Odds result	My model better than elo by	My model is better than odds by
2000	0.1932	0.2001	0.2017	3.60%	4.42%
2001	0.1975	0.2065	0.1996	4.55%	1.05%
2002	0.1914	0.1962	0.1974	2.54%	3.14%
2003	0.1985	0.2044	0.2081	3.00%	4.83%
2004	0.1863	0.1895	0.1930	1.72%	3.65%
2005	0.1951	0.1977	0.1980	1.34%	1.49%
2006	0.1898	0.1970	0.1996	3.77%	5.17%
2007	0.1877	0.1918	0.1900	2.23%	1.24%
2008	0.1872	0.1997	0.2035	6.63%	8.70%
2009	0.1848	0.1914	0.1929	3.59%	4.42%
2010	0.1945	0.2094	0.2149	7.62%	10.44%
2011	0.2017	0.2128	0.2194	5.47%	8.77%
2012	0.1832	0.1961	0.1980	7.01%	8.08%
2013	0.2044	0.2128	0.2076	4.08%	1.56%
2014	0.1893	0.2009	0.1991	6.11%	5.18%
2015	0.2061	0.2143	0.2162	3.99%	4.89%
2016	0.2079	0.2011	0.2041	-3.25%	-1.80%
2017	0.1975	0.2008	0.2077	1.65%	5.15%
2018	0.2009	0.2049	0.2080	2.01%	3.55%
2019	0.2074	0.2196	0.2137	5.88%	3.05%

Table 14 – improved Elo model performance in English league

The numbers represent an average mean absolute error of the predictions.

Combining the results of all championships into one table:

Championship	Number of seasons analysed	Average performance of my model compared to Elo model	Average performance of my model compared to odds
England	20	3.68%	4.35%
France	13	3.27%	3.22%
Spain	15	3.81%	4.52%
Germany	20	2.94%	3.10%
Italy	15	4.30%	4.82%
Portugal	3	4.40%	4.91%
Netherlands	2	4.08%	5.80%
Turkey	3	2.51%	3.12%

Table 15 – improved Elo model performance combined

The results of the experiment show that the predicting performance of the improved Elo model is consistently better than the Elo model and bookmaker odds adjusted to 0 to 1 number.

7. Machine learning

7.1. Definition

Machine learning is a branch of artificial intelligence focused on building applications that learn from data and improve their accuracy over time automatically [41]. An algorithm is a sequence of statistical processing steps. In machine learning, algorithms are 'trained' to find patterns and features in big amounts of data in order to make decisions and predictions based on new data. The better the algorithm, the more accurate the decisions and predictions will become as it processes more data.

Machine learning algorithms are often categorized as supervised, unsupervised and reinforcement [42]:

- Supervised learning: the computer is presented with example inputs and their desired outputs, and the goal is to learn a general rule that maps inputs to outputs. Supervised machine learning requires less training data than other machine learning methods and makes training easier because the results of the model can be compared to actual labelled results.

- Unsupervised learning: No labels are given to the learning algorithm, leaving it on its own to find structure in its input. Unsupervised learning can be a goal in itself, for example, discovering hidden patterns in data, or a means towards an end - feature learning. Unsupervised learning is less about automating decisions and predictions, and more about identifying patterns and relationships in data that humans would miss.
- Reinforcement learning: A computer program interacts with a dynamic environment in which it must perform a certain goal, for example, driving a vehicle or playing a game against an opponent. This model learns as it goes by using trial and error. A sequence of successful outcomes will be reinforced to develop the best recommendation or policy for a given problem.

There are four basic steps for building a machine learning model:

- Preparing a data set. The data set is divided into three subsets: the training set, the validation set, and the test set [43]. Common ratio used is 70% train, 15% validation, 15% test, but these numbers can vary. The ratio I used is 80% train, 10% validation, 10% test. The splitting is required as the process of determining the best model among several ones is training each model on the training set, evaluation of each model on the validation set, choosing the best on and evaluating it on the test set.
- Choosing an algorithm. The choice of an algorithm depends on several factors such as: size of the training data, accuracy and interpretability of the output, number of features and speed of the algorithm. It also depends on the type of the problem. Classification is the problem of predicting a discrete class label output, regression is the problem of predicting a continuous quantity output.
- Training the algorithm. Training the algorithm is an iterative process–it involves running variables through the algorithm, comparing the output with the results it should have produced, adjusting weights and biases within the algorithm that might yield a more accurate result, and running the variables again until the algorithm returns the correct result most of the time.
- Using and improving the model. The final step is to use the model with new data and, in the best case, for it to improve in accuracy and effectiveness over time.

7.2. Experiment

In order to create a machine learning model, it is necessary to provide it with historical data on football match results. The main problem is that the algorithm would learn on just team names and their results. But teams tend to change over time, they might become stronger or weaker due to transfers, aging or some other factors, also some teams are relegated and promoted between the leagues each year. Therefore, it was decided to bring my updated Elo rating model to replace team names with their Elo rating which they had before the match. So, the machine learning algorithm would learn not on team names but on Elo ratings. Such factors as form and motivation are also going to be used in the algorithm to increase the accuracy of the model.

There are two applicable methods of solving the problem: regression and classification. The regression is based on the goal difference, the classification is based on three options: 'home team wins', 'home team loses' and 'draw'.

The data used for the experiment is 5 championships (England, Germany, France, Italy and Spain), total of 84 seasons or 30440 matches.

According to the research conducted before the experiment, the best algorithms for regression are:

- Linear Regression - fits a linear model with coefficients w = (w1, …, wp) to minimize the residual sum of squares between the observed targets in the dataset, and the targets predicted by the linear approximation [44].
- Support Vector Machine (SVM) - constructs a hyperplane in multidimensional space to separate different classes. SVM generates optimal hyperplanes in an iterative manner, which is used to minimize an error. The core idea of SVM is to find a maximum marginal hyperplane that best divides the dataset into classes. The algorithm can be implemented for both regression and classification problems.
- Decision trees - non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. A tree can be seen as a piecewise constant approximation.

- Random Forest – it is based on ensemble learning which is a type of learning where different types of algorithms are joined or same algorithm multiple times to form a more powerful prediction model. The random forest algorithm combines multiple algorithms of the same type that are multiple decision trees, resulting in a forest of trees. The random forest algorithm can be used for both regression and classification tasks.

The best algorithms for classification are:

- Naive Bayes - Naive Bayes methods are a set of supervised learning algorithms based on applying Bayes' theorem with the "naive" assumption of conditional independence between every pair of features given the value of the class variable. Bayes' theorem states the following relationship, given class variable y and dependent feature vector x_1 through x_n:

  $$P(y|x_1, \dots, x_n) = \frac{P(y)P(x_1, \dots, x_n|y)}{P(x_1, \dots, x_n)}$$

- Bagging - it is an ensemble meta-estimator that fits base classifiers each on random subsets of the original dataset and then aggregate their individual predictions to form a final prediction. Such a meta-estimator can typically be used to reduce the variance of a black-box estimator (for example, a decision tree), by introducing randomization into its construction procedure and then making ensemble out of it.
- Random Forest
- Decision Trees

The regression algorithms are going to be evaluated with the help of the following metrics:

- R^2 or coefficient of determination - compares the model's predictions to the mean of the targets. Values can range from negative infinity to 1. For example, if all the model does is predict the mean of the targets, its R^2 value would be 0. And if the model perfectly predicts a range of numbers it's R^2 value would be 1.
- Mean absolute error (MAE) - The average of the absolute differences between predictions and actual values. It gives an idea of how wrong the predictions are.

- Mean squared error (MSE) - The average squared differences between predictions and actual values. Squaring the errors removes negative errors. It also amplifies outliers - samples which have larger errors)

The classification algorithms are going to be evaluated with the help of the following metrics:

- Accuracy: tells how often the algorithms predicts the output correctly. The accuracy is given in decimal form. Perfect accuracy is equal to 1.0, which is the same as getting the prediction right 100% of the time.
- F1 score: combination of the recall (the proportion of actual positives which were correctly classified) and the precision (the proportion of positive identifications which were actually correct)

7.3. Results

Baseline models of chosen algorithms give the following results:

Model	R^2 score	MAE	MSE
Decision Tree	-0.51310	0.26112	0.11083
Linear Regression	0.20927	0.19198	0.05626
Random Forest	0.08429	0.20532	0.06536
SVM	0.20636	0.19214	0.05629

Table 16 – Baseline models regression results

Model	Accuracy	F1
Bagging	0.46342	0.45689
Decision Tree	0.42053	0.41355
Naive Bayes	0.51632	0.44755
Random Forest	0.47066	0.45342

Table 17 – Baseline models classification results

The next step is tuning hyperparameters in the models. As different models have different parameters, it affects the performance of the models. The tuning can be performed by simple selection method, by randomized search or the grid search. First of all, the grid of several values for several parameters is chosen for a model, for example:

```
grid = {"n_estimators": [10, 100, 200, 500, 1000, 1200],
        "max_depth": [None, 5, 10, 20, 30],
        "max_features": ["auto", "sqrt"],
        "min_samples_split": [2, 4, 6],
        "min_samples_leaf": [1, 2, 4]}
```

As model testing for each of the combinations takes up to several days, random search gives information which of the n randomly chosen combinations gives the best results. After receiving the best of the random combination, grid search is used – it searches across a grid of hyperparameters exhaustively. The numbers in the new grid are chosen according to the parameters obtained from the previous search.

After tuning of hyperparameters, the results were as follows:

Model	R^2 score	MAE	MSE
Decision Tree	0.13497	0.18993	0.06019
Linear Regression	0.19164	0.19198	0.05626
Random Forest	0.17056	0.17077	0.05775
SVM	0.20636	0.19214	0.05629

Table 18 – Tuned models regression results

Model	Accuracy	F1
Bagging	0.51959	0.44202
Decision Tree	0.52724	0.44304
Naive Bayes	0.53434	0.44375
Random Forest	0.52987	0.43784

Table 19 – Tuned models classification results

After receiving the results, the models are saved using the pickle python library.

All the results tend to reach approximately the same value, the reason for which might be the randomness or luck which was explored in the Introduction of the thesis. It can be assumed that even the best possible models would not be able to achieve the results much better than those in my work due to the factor of luck.

8. Poisson

8.1. Definition

Another model that is able to predict football results is Poisson distribution. This probability distribution was first introduced by Simeon-Denis Poisson in 1838. It is a discrete probability distribution and it aims to express the probability of occurrence of certain events in a fixed interval. It is assumed that the mean number of events occurring during the time interval is known and that the time difference between any event and the event that follows it is independent of the previous time differences [45]. While the Poisson distribution is applied to problems with certain fixed time unit's interval, it can also be successfully applied to football match results.

The general focus of the Poisson distribution is a variable event which occurs at a certain time interval and the number of events observed in this range is considered to be a random variable for the Poisson distribution. The expected value of the number of events occurring in this fixed range (the mean number of occurrences) is fixed as λ, and this mean value is proportional to the range length.

The probability of occurrence of a k-number (k = 0, 1, 2, 3 ...) of any non-negative phenomenon is expressed as follows:

$$f(k,\lambda) = \frac{\lambda^k e^{-\lambda}}{k!} \tag{20}$$

where: e is the base of the natural logarithm, k is the number of occurrences of a variable event which probability is given with a function, λ is the expected value of occurrence of an event in a given fixed interval.

8.2. Experiment

In order to calculate goal expectancy for a home team in a match (HGE), the following formula is used:

$$HGE = HAS * ADS * AGH \quad (21)$$

Where: HAS is home team's home attacking strength (calculated as a home team's average goals scored per game at home divided by average goals scored per game at home in the league), ADS is away team's away defensive strength (calculated as an away team's average goals conceded per game away divided by average goals conceded per game away in the league) and AGH is average goals scored home per game in the league. Accordingly, away team goal expectancy (AGE) is calculated by the formula:

$$AGE = AAS * HDS * AGA \quad (22)$$

Where: AAS is the away team's away attacking strength, HDS is the home team's home defensive strength and AGA is average goals scored away per game in the league.

It is obvious that in order to calculate these numbers, it is necessary to have some historical data, but due to constant changes in leagues, such as teams relegating and promoting, it is only possible to calculate the data within only one season. Therefore, the model can only predict the second half of the season after 'learning' on the data from the first half of the season, continuing to improve after the mid-season as the new data is getting available.

The procedure can be seen in the scripts located in the 'poisson' folder in the attachments. The outcomes of the model are provided in two forms: the number between 0 and 1 for the model to be comparable with the Elo models which takes into account the goal difference, and in the form of the win/draw/loss outcome.

8.3. Results

The results for the Poisson model are: 0.228 mean absolute error for the first form and 0.525 accuracy for the second form.

The results appeared to be not the best among other analysed models. The reason might be that the model only considers goals scored and goals conceded by the teams, it does not

take into account teams' strengths. Let's consider a case, when team A lost 7 games 0:1 and won only one game 7:0, because of, for example, a red card for an opponent team, and team B won 7 games 1:0 but lost only one game 0:7 due to some unexpected reasons. Model will see those teams equally, because they would have the same goal difference, where it is obvious that one is much worse than the other. Predicting draws is also difficult for the model, as usually only about 25% of matches are draws, therefore, the model mostly predicts the win of one of the teams playing. For example, for 20 seasons of English Premier League it has predicted only 48 draws, where actually the number was 1914.

9. Conclusion

Six different factors affecting football match results were discussed and evaluated in this thesis. The Elo ranking system was also discussed and the factors were introduced into the model. The improved Elo model was created and as it was found that it predicts football match results better than the original Elo model for 2.9%, it might be also better for the use of ranking teams.

The results of the conducted experiments show that the best model is the machine learning algorithm based on my updated Elo model. The best result for predicting a football game outcome is Naïve Bayes algorithms that gave 0.535 accuracy, in other words, it predicts 53.5% of the games correctly. Comparing this result with the other machine learning predicting football match results study [46], my model is 2.4% better, due to the improved Elo model. Unfortunately, my model cannot be implemented on the football matches since 2019 till now due to the Covid-19 and the effects it had on football, for example, the home advantage has decreased significantly, as the fans are restricted to support the teams on the stadiums [47].

9.1. Future work

All the models could be improved by providing more historical data as well as deeper data which is not possible to find in open sources of information. It would add or improve other factors that affect the football match results, therefore, improving the updated Elo formula giving better predictions and rankings. Also, international matches can be implemented into the Elo model, this might potentially increase predictability of the model.

9.2. Probable uses

The rating described in this work may be useful for national and international cups. As currently most of the cup fixtures are decided randomly via closed draw, it may cause best teams playing against each other at earliest stages decreasing overall teams quality at later stages, therefore, decreasing fans interest in the cup causing losing money from TV-broadcast and matchday earnings.

Low team quality in later stages with random draw is not the only problem. Even when top teams play against each other at earlier stages it does not get as much attention as it gets even when these teams play regular season games in the championship. There are several examples in English FA Cup presented below:

Team 1	Team 2	Capacity	Cup stage	Cup attendance	League attendance	Season
Wolverhampton	Liverpool	32050	Round of 64	25849	31358	2018/2019
Arsenal	Man Utd	60260	Round of 32	59571	60000	2018/2019
Man City	Burnley	55097	Round of 64	53356	54118	2017/2018
Liverpool	Everton	53394	Round of 64	52513	53082	2017/2018
Southampton	Arsenal	32505	Round of 32	31288	31474	2016/2017
Everton	Leicester	39573	Round of 64	35493	39573	2016/2017

Table 20 – Attendance at league matches and early cup matches

The table shows that the matches between top-level teams in the earliest cup stages do not get as much attention as matches between these exact teams on the same field during one season but in championship.

The example of the working similar model is Grand Slam tennis tournaments [48], where 32 top players are seeded in the way that they will not meet each other at the earliest stages.

For instance, the best player cannot meet the second-best player before the finals and cannot meet third and fourth best players before the semi-finals. That method is applied only for top-32 players of the given tournament. The other 96 (there are 128 participants in the Grand Slam tournaments) players are randomly distributed via a live public draw.

The offered method can also help to avoid the cases when one of the top teams get an easier draw comparing to other top teams resulting in less competitive final stages of the cup bringing down the interest in the championship.

One of the advantages may be an increased number of matches between top teams and teams from bottom leagues (for example, English FA Cup where games between top teams and lower league teams are not rare), therefore, the number of competitive games would increase.

One of the disadvantages of this method for tennis is that early stages are not as competitive as it would have been with a random draw. For professional football it is not necessarily a case, because early cup stages are usually held during first months of the season where top teams might have some extra fixtures in international competitions. Therefore, cup matches against weaker teams would not strongly affect the performance in the international competitions, as top teams prefer letting players from the reserve or young squad to play in the earliest stages of the cups.

One more possible use of the model could be the fixtures schedule. According to the interview of Glenn Thompson [49], who is in charge of scheduling English Premier League season games there is no particular algorithm to prevent clubs to have several games in a row against strong opponents, therefore, my model could help football associations to avoid such cases when one team plays several games against stronger opponents in a row. By doing that, the competitiveness of the league might increase, therefore, increasing viewers interest and income.

References

[1] Laws of the Game, 2015/2016. [online]. Accessed April 2021. Available from: https://img.fifa.com/image/upload/datdz0pms85gbnqy4j3k.pdf

[2] Michael J. Mauboussin. “The Success Equation”. Harvard Business Review Press Boston, Massachusetts, USA.

[3] Betting Odds explained: How are football odds calculated. [online]. Accessed April 2021. Available from: https://footballwhispers.com/blog/betting-odds-explained/

[4] About USCF. [online]. Accessed April 2021. Available from: http://www.uschess.org/index.php/About-USCF/

[5] James Grime. (2018, February 15). The Elo Rating System for Chess and Beyond. [Video]. Youtube. Accessed April 2021. Available from: https://www.youtube.com/watch?v=AsYfbmp0To0

[6] Elo rating system. [online]. Accessed April 2021. Available from: https://www.chess.com/terms/elo-rating-chess

[7] Men’s Ranking by FIFA. [online]. Accessed April 2021. Available from: https://www.fifa.com/fifa-world-ranking/ranking-table/men/

[8] “Ties in paired-comparison experiments: A generalization of the Bradley-Terry model”, Journal of the American Statistical Association, 62, pp. 194-204

[9] “The Glicko System”. By Prof. Mark E. Glickman. Available at http://www.glicko.net/research/gdescrip.pdf

[10] J. Lasek, Z. Szlavik and S. Bhulai. ”The predictive power of ranking systems in association football”. Available at http://lasek.rexamine.com/football_rankings.pdf

[11] Wilson, Jonathan. Is three points for a win good for football? [online]. Accessed April 2021. Available from: https://www.theguardian.com/sport/blog/2009/feb/05/question-jonathan-wilson-three-points

[12] Garciano. Luis. 2005. Sabotage in Tournaments: Making the Beautiful Game a Bit Less Beautiful. Available from:

https://www.researchgate.net/publication/4842950_Sabotage_in_Tournaments_Making_the_Beautiful_Game_a_Bit_Less_Beautiful

[13] The BMJ. Correlation and Regression. [online]. Accessed April 2021. Available from: https://www.bmj.com/about-bmj/resources-readers/publications/statistics-square-one/11-correlation-and-regression

[14] Statistics How To. Regression Analysis. [online]. Accessed April 2021. Available from: https://www.statisticshowto.com/probability-and-statistics/regression-analysis/

[15] Bundesliga. 2018. xG stats explained: the science behind Sportec Solutions' Expected goals model. [online]. Accessed April 2021. Available from: https://www.bundesliga.com/en/bundesliga/news/expected-goals-xg-model-what-is-it-and-why-is-it-useful-sportec-solutions-3177

[16] Kilday. James. 2018. Re-Introducing Expected Points – (XP/XPoints). [online]. Accessed April 2021. Available from: https://www.modernfitba.com/blogs/2018/9/8/re-introducing-expected-points-xpxpoints

[17] IBM Cloud Education. 2020. Monte Carlo Simulation. [online]. Accessed April 2021. Available from: https://www.ibm.com/cloud/learn/monte-carlo-simulation

[18] World Football Elo Ratings. [online]. Accessed April 2021. Available from: https://www.eloratings.net

[19] Magee. Will. 2017. How The 2002 World Cup Became the Most Controversial Tournament in Recent Memory. [online]. Accessed April 2021. Available from: https://www.vice.com/en/article/ywgx4y/how-the-2002-world-cup-became-the-most-controversial-tournament-in-recent-memory

[20] Mexico City. Britannica. [online]. Accessed April 2021. Available from: https://www.britannica.com/place/Mexico-City

[21] Rumsby. Ben. 2018. Russia leave World Cup rivals in their wake to top running stats - what's behind the hosts' revival? [online]. Accessed April 2021. Available from: https://www.telegraph.co.uk/world-cup/2018/06/20/russia-leave-world-cup-rivals-wake-top-running-stats/

[22] Widermann, Diana; Barton, Robert A.; Hill, Russel A. (2011). "Evolutionary perspectives on sport and competition". In Roberts, S. Craig (ed.). Applied Evolutionary Psychology. Oxford University Press.

[23] Doyle. Paul. 2013. Why is Nigeria the world's toughest league in which to get an away win? [online]. Accessed April 2021. Available from: https://www.theguardian.com/football/blog/2013/oct/29/nigeria-toughest-league-win-away

[24] Picture: Province of Las Palmas. Wikipedia. [online]. Accessed April 2021. Available from: https://en.wikipedia.org/wiki/Province_of_Las_Palmas#/media/File:Las_Palmas_in_Spain_(real_location).svg

[25] Picture: Madeira. Wikipedia. [online]. Accessed April 2021. Available from: https://cs.wikipedia.org/wiki/Madeira_(ostrov)#/media/Soubor:Madeira_in_its_region.svg

[26] Villegas Gama. Karla. 2013. Loudest Stadiums in World Football. [online]. Accessed April 2021. Available from: https://bleacherreport.com/articles/1821923-loudest-stadiums-in-world-football

[27] Lago Penas. Carlos. 2019. The Influence of Age on Footballers' Performance. [online]. Accessed April 2021. Available from: https://barcainnovationhub.com/the-influence-of-age-on-footballers-performance/

[28] Sporting Intelligence. 2009. "Form" in Football. [online]. Accessed April 2021. Available from: https://www.sportingintelligence.com/2009/12/16/fact-or-fiction-myths-in-football/

[29] "Hot Hand" phenomenon. [online]. Accessed April 2021. Available from: https://www.sports.ru/tribuna/blogs/bankshot/2877553.html?sl=1

[30] Miller, J. B., & Sanjurjo, A. (2018). Surprised by the hot hand fallacy? A truth in the law of small numbers. Econometrica, Vol. 86, No.6, pp. 2019–2047. Available from: https://poseidon01.ssrn.com/delivery.php?ID=288029089119125072068078018001025091103043056088031004087020093093071124089007103068017098101006051012034021074007122070023014122090028033029000093127010101077103015084017084101115003106031101068094125099083077081097099084066095080104118031030122105001&EXT=pdf&INDEX=TRUE

[31] Valenzuela. Jose. 2020. Player's Motivation in Tough Situations. [online]. Accessed April 2021. Available from: https://barcainnovationhub.com/players-motivation-in-tough-situations/

[32] The Stats Zone. 2017. Is High Ball Possession Key To Success In European Football? [online]. Accessed April 2021. Available from: https://www.thestatszone.com/archive/is-ball-possession-key-success-european-football-13776

[33] Soccerment Research. 2021. Ball Possession in European Football. [online]. Accessed April 2021. Available from: https://soccerment.com/ball-possession-european-football-part-1/

[34] Goal. 2019. What is Red Card in Football. [online]. Accessed April 2021. Available from: https://www.goal.com/en-us/news/what-is-a-red-card-in-soccer/1j1nq8o7xwart1vujfq7kpbq0j

[35] Transfer. Wikipedia. [online]. Accessed April 2021. Available from: https://en.wikipedia.org/wiki/Transfer_(association_football)

[36] Brand. Gerard. 2015. How the Bosman rule changed football – 20 years on. [online]. Accessed April 2021. Available from: https://www.skysports.com/football/news/11096/10100134/how-the-bosman-rule-changed-football-20-years-on

[37] Athletic Interest. (2021, January 4). Why the Ronaldo Deal Was A Winning Bet [Video]. Youtube. Accessed April 2021. Available from: https://www.youtube.com/watch?v=kIRBW8EP6R4

[38] Simon Kuper, Stefan Szymanski. “Soccernomics”. Nation Books, A Member of the Perseus Books Group. 2014.

[39] Veth. Manuel. 2020. Corruption charges against Lim & Mendes - Prosecutors used TM market values. [online]. Accessed April 2021. Available from: https://www.transfermarkt.com/corruption-charges-against-lim-amp-mendes-prosecutors-used-tm-market-values/view/news/374820

[40] Transfermarkt Scout Interview. [online]. Accessed April 2021. Available from: https://www.sports.ru/tribuna/blogs/allozdravstvuite/2879249.html?sl=1

[41] IBM Cloud Education. 2020. Machine Learning. [online]. Accessed April 2021. Available from: https://www.ibm.com/cloud/learn/machine-learning

[42] Machine Learning. Sas. [online]. Accessed April 2021. Available from: https://www.sas.com/en_us/insights/analytics/machine-learning.html

[43] Draelos. Rachel. 2019. Best Use of Train/Val/Test Splits. [online]. Accessed April 2021. Available from: https://glassboxmedicine.com/2019/09/15/best-use-of-train-val-test-splits-with-tips-for-medical-data/

[44] Choosing the right estimator. [online]. Accessed April 2021. Available from: https://scikit-learn.org/stable/tutorial/machine_learning_map/index.html

[45] Inan. Tugbay. 2020. Using Poisson model for goal prediction in

European football. [online]. Accessed April 2021. Available from: https://rua.ua.es/dspace/bitstream/10045/107443/1/JHSE_16-4_InPress_16.pdf

[46] Herbinet. Corentin. 2018. Predicting Football Results Using Machine Learning Techniques. [online]. Accessed April 2021. Available from: https://www.imperial.ac.uk/media/imperial-college/faculty-of-engineering/computing/public/1718-ug-projects/Corentin-Herbinet-Using-Machine-Learning-techniques-to-predict-the-outcome-of-profressional-football-matches.pdf

[47] StatsbyLopez. 2016. On soccer's declining home field advantage. [online]. Accessed April 2021. Available from: https://statsbylopez.com/2016/05/13/on-soccers-declining-home-field-advantage/

[48] Amenechi. Akosa. 2017. 32 Not Out: How Does the Number of Seeds Affect a Grand Slam Draw? [online]. Accessed April 2021. Available from: https://www.perfect-tennis.com/32-not-out-how-does-the-number-of-seeds-affect-a-grand-slam-draw/

[49] Vincent. Bobby. 2020. How the Premier League fixture list is decided. [online]. Accessed April 2021. Available from: https://www.manchestereveningnews.co.uk/sport/football/football-news/how-premier-league-fixture-list-18783350