**1. Introduction**

Can the game of football, complex and unpredictable by nature, be understood better with the help of artificial intelligence and machine learning?

As one of the most famous sports, football is played by approximately 250 million players in over 200 countries and has billions of fans worldwide. There are three possible outcomes for a game: win, lose, or draw. The margins for error and possible outcomes in each match are relatively small; coupled with the myriad other types of influence, from individual player abilities and current form, to more external factors such as weather, or even the psychological influence of fans in the stadium [1], we can agree predicting the outcomes of certain matches, scores and league table positions can reach dizzying levels of complexity.

There are many metrics to be measured during a football game, and part of the art will be choosing which inputs will have a positive impact on the systems predictions. What weighting does the number of corners taken by a team have on the final score? How about yellow cards? Possession statistics?

Our data sets will be taken from the top tier of English football, the Premier League, with the features we are most likely to use including:

* Final score
* Number of shots off target
* Number of shots on target
* Possession
* Number of passes
* Passing accuracy
* Yellow cards
* Red cards
* Home and Away
* Corners
* Fouls

The challenge for our system will be to analyse the noise each metric produces in relation to the final outcome, and then through learning determine a weighting for each input that will result in more accurate predictions.

Therefore, we will evaluate our system based on how accurate our predictions were based on historical data. As we can’t see into the future, it makes sense to train our system on a large portion of the season and the test it on the remaining matches. We will then do a final test of the system on matches from the second half of the 2020/2021 season up until the project deadline and produce a final evaluation of the system’s accuracy.

An extension to our evaluation of the system would be comparing our predictions to the spread bets from betting companies, i.e. is our system better at predicting outcomes than following the wisdom of the crowd?

**2. Literature Review**

Football is a game with few goals, which means that many inconspicuous factors can determine the final result. Therefore, it is quite difficult to choose features and classification methods that help predict. In the paper [2], authors list the steps of the system design to build the prediction model.

**2.1 Feature Selection**

Many factors such as the form of the football team, home advantage, the overall quality of the team, and the personal qualities of the players in the team will affect the final result of the football game. The problem of selecting features can be solved in two ways [3]. The first way implies that we have a certain understanding of the problem, and based on this knowledge, we choose those characteristics that we think will most affect the final result. The other way means that we have little or no knowledge of the problem, so we select all the features that may affect the results, and then gradually try to determine the most influential features[4].

**2.2 Selecting Learning Algorithms**

In order to achieve better prediction results, various algorithms are considered to determine which provides better results.

***Bayesian networks:***

Bayesian networks, BNs, provide a means for capturing, displaying, and making available in a usable form knowledge, often obtained from experts in a given field[5].

When approaching a new problem there are two techniques which are commonly used. The first assumes we have some idea how the situation under investigation works, construct a model, and using this model select the attributes believed to contribute to the effect under investigation. An example of this approach to this type of problem is given. The second approach assumes little knowledge of the underlying mechanisms involved so we look at all the probably relevant attributes and try to determine those which have the most significant effect. This is still in effect the construction of an a priori model, but only a very informal one[6].

***LogitBoost:***

The ‘*LogitBoost’* algorithm is a boosting algorithm. The Boosting algorithm consists of a set of weak classifiers. If they are used as independent classifiers, their performance will be poor, but the combination will work well.

***K-nearest neighbour algorithm:***

It is the representative algorithm from the group of lazy classifiers. Classification of new examples is done by finding k nearest neighbors in the space of features (typically using Euclidean distance) from the examples in the learning set. Based on these examples voting or another method determines classification of new examples[7].

**2.3 Conclusion**

Predicting the results of football matches poses an interesting challenge due to the fact that the sport is so popular and widespread. However, predicting the outcomes is also a difficult problem because of the number of factors which must be taken into account that cannot be quantitatively valued or modelled. As part of this work, a software solution has been developed in order to try and solve this problem. During the development of the system, a number of tests have been carried out in order to determine the optimal combination of features and classifiers. There is room for further improvement, primarily in the area of feature selection. If we were to model the form for each and every player in the match, we could probably achieve better results. This way we could monitor each player's form during the season and determine its influence on the final score. Besides, larger data set for learning would also help to predict future outcomes[8].

**3. Method**

Since this is a competitive market, we are fortunate enough to have many datasets at our disposal. As mentioned previously, part of our analysis will centre around which inputs are most important, as too many may impact our running time and memory usage.

The method we will be using is k-fold cross validation, over the last decade of Premier League seasons, and then finally evaluating our model on the current season. If we are to split each season by a ratio of 9:1, we will use the first 340 matches as our training dataset, and the subsequent 40 matches as our testing dataset. This method will help us avoid overfitting the dataset.

Since our modelling goal is to predict the outcome of matches, we will need to use a multinomial classifier with ‘win’, ‘lose’ or ‘draw’ as labels for our dataset. Numerically, and perhaps to help us when calculating our loss, we could represent these outcomes as 1, -1 and 0 respectively.

We will use a neural network and begin with arbitrary weights for each of the features, normalising them to avoid any issues with scale, say in the interval [-1,1]. As we train our algorithm, these weights will be what our machine learns on and tries to improve. Propagation will be an important guide for us: if our label says win but we predicted a loss, the machine must change its weighted values to minimise risk.

Once the machine has predicted a set of outcomes, we will use binary classification to classify which results were correct and which were incorrect.

**4. Initial Experimentation and Analysis**

**5. Discussion**

**5.1 Current Methods**

It’s no secret that sports betting companies use incredibly complex systems and employ sophisticated analytical techniques to make sure they stay one step ahead.

The machine learning techniques they use will be vast, most likely inputting data that the public can’t gain access to, or at the very least using techniques and processing power that is too expensive for the general public to access.

One such example of this is image recognition and mapping software now currently used widely in the industry. Created by in house data teams from football teams themselves to provide better opportunities for post-game analysis and tactic development, it wasn’t long before betting companies started to invest in similar programs to give them an added advantage.

Football has the power to take us completely by surprise (Leicester’s 5000/1 title winning season in 2016), particularly when analysing individual match outcomes, but when seen from a broader perspective some metrics are surprisingly easy to predict. For instance, we can model the expected total number of goals in a game as a Poisson distribution [9]. This method of course has its limitations (the only factor the distribution considers is the result), but it’s interesting to see how suddenly predictable the game can be.

**5.2 Ethics**

Is it right to use a machine learning algorithm to make better predictions on sports results? One could argue computing power should be left out of this sphere, at the expense of good old fashioned gut feeling. We wouldn’t think it fair for a poker player to elicit computing power to give themselves an edge, so is it the same for sports betting?

Arguably not, particularly when the player most bettors are up against is not only the unpredictable nature of the game, but sports betting companies with the financial and computing power to place themselves at a great advantage. Is their use of these programs unethical, particularly when it’s used to stack the odds against the betting community’s finances, not to mention the betting industry’s unsatisfactory work in tackling gambling addiction? Perhaps a portion of the profits generated from these powerful machines could be used to help tackle this growing problem.

**6. Work for TB2**

The first half of the term will centre around writing or adapting an existing algorithm to test on our dataset. Much work can be done on striking the perfect balance between memory usage and making sure the machine is sophisticated enough to produce competitive results.

Once we happy with the complexity of our system and satisfied we have exhausted its potential, one further extension to evaluating its success could be matching it up against the bookmaker’s odds and see if the machine would be capable of providing a positive financial return.

**7. Other Project Ideas**

***Picture recognition***

With the development of search engines, people are more reliant on the help of the program to find out the content they need, not only words, but also pictures. We can distinguish the cat pictures and dog pictures, but computers recognise things differently. We need to teach them with lots of pictures so that they can figure out the difference. But when the algorithm learns enough data and well programmes, it can do things that humans cannot, such as the texture or position of the picture.

***Language Translation***

With the development of globalization, there is more communication within people from different countries. The huge difference between cultures leads to the problem which is the improper translation. People can normally recognise sentences translated by translators because of the different structures and choice of words. We think that it can be improved with the addition of filters or tags. From our perspective, if we can add a tag system to stuff we want to translate, then train the algorithm with more exact context, then it should come out with more authentic results.

***Virtual assistant***

Siri, Alexa, Cortana, guess we are familiar with at least one of them. Setting alarm clock, counting time, making a call, virtual assistant follow our command by recognizing our voice and certain words. We are interested to find out the way of how voice can be recognized.

**8. Discussion of Topics**

***Machine Learning***

As the chosen method of the project, we have explained a lot about it. It is the core of artificial intelligence and the fundamental way to make computers intelligent. The research directions of traditional machine learning mainly include research on decision trees, random forests, artificial neural networks, and Bayesian learning.

***Decision Trees***

Decision trees are based on the known probability of occurrence of various situations, by constructing a decision tree to obtain the probability that the expected value of the net present value is greater than or equal to zero, evaluate project risk, and judge its feasibility. A decision tree is a tree structure in which each internal node represents a test on an attribute, each branch represents a test output, and each leaf node represents a category. In our project, it can be used for linking between individual factors, such as the strength of the offense of a player can lead to a chance of the goal.

***Markov Decision Processes***

MDP is used to simulate the achievable randomness strategy and rewards of the agent in an environment where the system state has Markov properties. MDP is a mathematical model that simulates the random policy and rewards of an agent in the environment, and the state of the environment has Markov properties. It contains 5 model elements, state, action, policy, reward and return. In our project, it can be used to simulate the strategy of a team in different situations. But due to the complexity of the reinforcement learning we may not use this method in the project.

***Game Theory***

Strictly speaking, game theory is mainly a mathematical model of conflict and cooperation between rational decision makers. From my perspective, it is a study of how to play various games in this world reasonably. Traditional machine learning is mostly regarded as an optimization problem. What we need to do is to find an algorithm that can search for the optimal solution. At this time, the idea of game theory is very important. Generalization is sometimes referred to as "structural risk minimization."

**9. References**

[1] J. Hucaljuk and A. Rakipović, "Predicting football scores using machine learning techniques," 2011 Proceedings of the 34th International Convention MIPRO, Opatija, 2011, pp. 1623-1627.

[2] Balla, Radha-Krishna. "Soccer match result prediction using neural networks." (2007).

[3] A. Joseph, N.E. Fenton and M. Neil, "Predicting football results using Bayesian nets and other machine learning techniques", Jelsevier B.V., 2006.

[4] Tom M. Mitchell, "Machine Learning" in , McGraw-Hill Science/Engineering/Math, 1997.

[5] L. Breiman, "MACHINE LEARNING - Random Forests" in , Kluver Academic Publishers, vol. 45, pp. 5-32, 2001.

[6] B. DalbeloBasic, M. Cupic and J. Snajder, Umjetne neuronske mreze notes for Artificial intelligence course, FER Zagreb, 2008.

[7] Michie, Donald, David J. Spiegelhalter, and C. C. Taylor. "Machine learning." Neural and Statistical Classification 13.1994 (1994): 1-298.

[8] Russell, Stuart, and Peter Norvig. "Artificial intelligence: a modern approach." (2002).

[9]  https://www.prosoccer.eu/betting-theory/poisson-distribution-football/