MatchPredictor Final Report

Atom Arce, Spencer Ball, Markus Kunej

ECE324: Introduction to Machine Intelligence
December 6, 2020
Word Count: 1944 (0% penalty)

Table of Contents

Introduction

Goal

To create a neural network which predicts the outcome of a soccer match (home team winning, away team winning, or a draw) for the top leagues in Europe.

Why?

Soccer is the most popular sport in the world, with an estimated 4 billion fans [1]. A more accurate match prediction tool could be used by these fans to learn how their favourite teams will perform, as well as attract new fans by providing insight into the sport. Also, soccer accounts for the largest volume of the $250 billion sports betting market [3]. A more accurate predictor would be highly sought-after by these companies, potentially leading to great monetary gain.

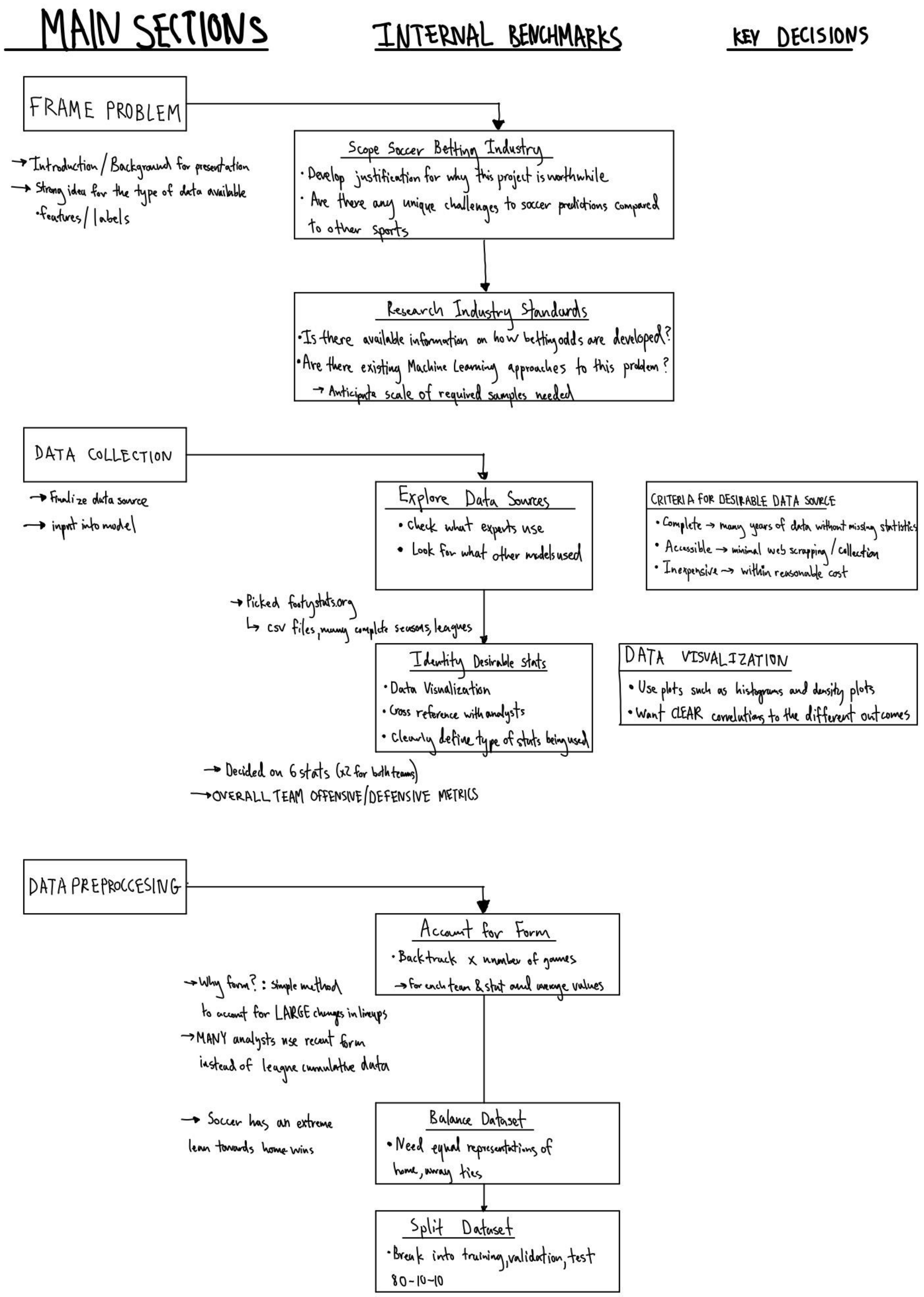
A Difficult Problem With Much Uncertainty

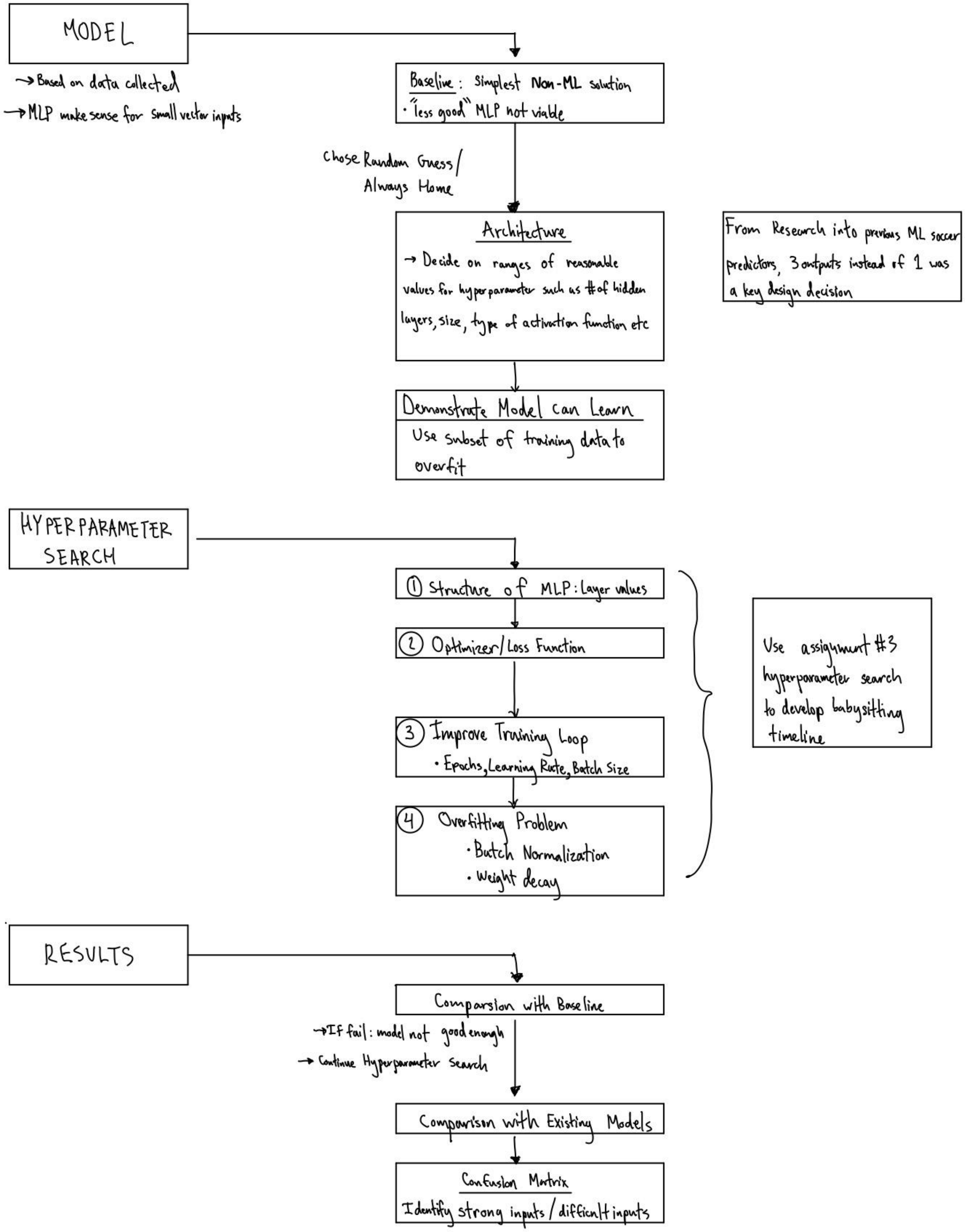
On the surface this may seem like a simple problem to solve -- there are only three possible outcomes and often one team is heavily favoured over the other -- it is a far more difficult one because of soccer's natural uncertainty.

As a testament to how unpredictable soccer can be, possibly the greatest upset in sports history occurred in 2016 when Leicester FC won the English Premier League, defying 1/66000 betting odds set at the beginning of the season [4].

Currently, human experts successfully predict only 50-55% of the matches [4], while past attempts at using neural networks yielded accuracies between 48% and 55% [5]. We believe there is more room for improvement using neural networks, considering the vast amount of data available on both teams before every match. Combining this with our team's passion for soccer is why we were motivated to tackle this problem.

Illustration/Diagram

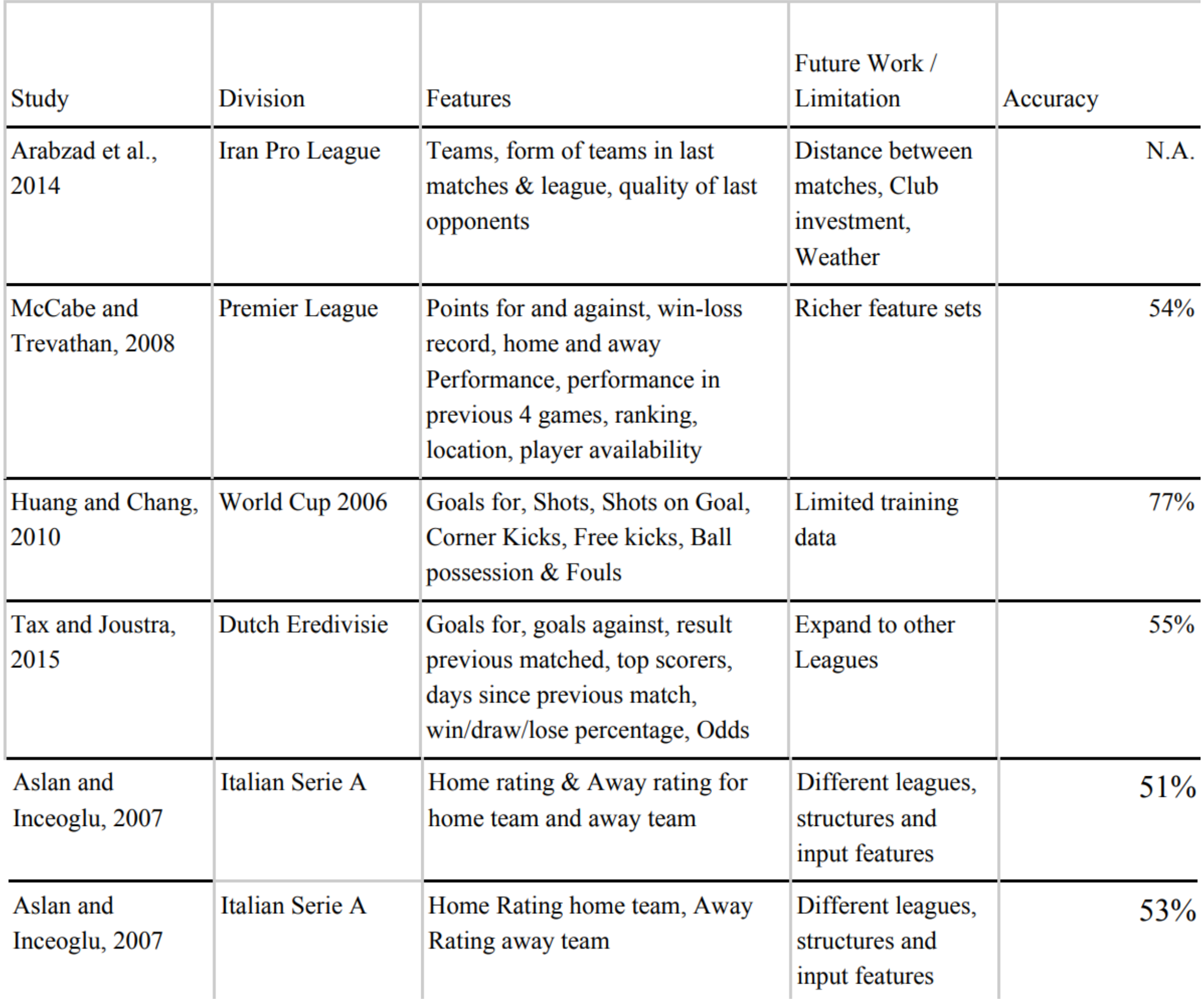

MAIN SECTIONS
INTERNAL BENCHMARKS
KEY DECISIONS
FRAME PROBLEM
→ Introduction / Background for presentation
→ Strong idea for the type of data available
•features / labels
Scope Soccer Betting Industry
• Develop justification for why this project is worthwhile
• Are there any unique challenges to soccer predictions compared to other sports
Research Industry Standards
•Is there available information on how betting odds are developed?
•Are there existing Machine Learning approaches to this problem?
→ Anticipate scale of required samples needed
DATA COLLECTION
→ Finalize data source
→ input into model
Explore Data Sources
• Check what experts use
• Look for what other models used
CRITERIA FOR DESIRABLE DATA SOURCE
• Complete → many years of data without missing statistics
• Accessible → minimal web scrapping / collection
• Inexpensive → within reasonable cost
→ Picked footystats.org
↳ csv files, many complete seasons, leagues
Identity Desirable stats
• Data Visualization
• Cross reference with analysts
• Clearly define type of stats being used
DATA VISUALIZATION
• Use plots such as histograms and density plots
• Want CLEAR correlutions to the different outcomes
→ Decided on 6 stats (x2 for both teams)
→ OVERALL TEAM OFFENSIVE/DEFENSIVE METRICS
DATA PREPROCCESING
Account for Form
• Backtrack x number of games
→ For each team & stat and average values
→ Why form? : simple method to account for LARGE changes in lineups
→ MANY analysts use recent form instead of league cumulative data
→ Soccer has an extreme lean towards home wins
Balance Dataset
• Need equal representations of home, away ties
Split Dataset
• Break into training, validation, test
80-10-10


MODEL
→ Based on data collected
→ MLP make sense for small vector inputs
Baseline: Simplest Non-ML solution
· "less good" MLP not viable
Chose Random Guess / Always Home
Architecture
→ Decide on ranges of reasonable values for hyperparameter such as # of hidden layers, size, type of activation function etc
From Research into previous ML soccer predictors, 3 outputs instead of 1 was a key design decision
Demonstrate Model can Learn
Use subset of training data to overfit
HYPERPARAMETER SEARCH
① Structure of MLP: Layer values
② Optimizer/Loss Function
③ Improve Training Loop
· Epochs, Learning Rate, Batch Size
④ Overfitting Problem
· Batch Normalization
· Weight decay
Use assignment #3 hyperparameter search to develop babysitting timeline
RESULTS
Comparsion with Baseline
→ If fail: model not good enough
→ Continue Hyperparameter Search
Comparison with Existing Models
Confusion Matrix
Identify strong inputs / difficult inputs

Background & Related Work

There have been several related works in this field, but we will be looking closely at the 2019 paper *“Football Result Prediction by Deep Learning Algorithms”* by Stefan Samba. In it, Samba explores previous works, while also creating his own neural network to predict soccer match outcomes.

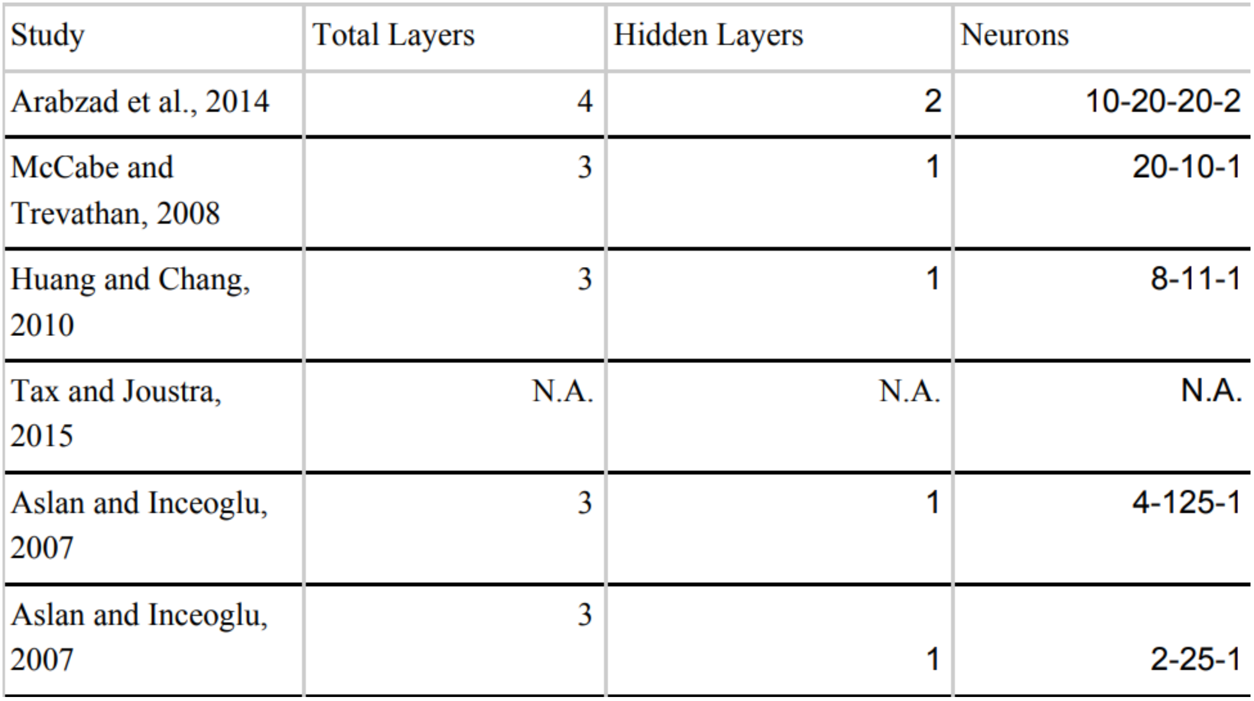
Table 1: An overview of studies reviewed by Samba [5]



Study	Division	Features	Future Work / Limitation	Accuracy
Arabzad et al., 2014	Iran Pro League	Teams, form of teams in last matches & league, quality of last opponents	Distance between matches, Club investment, Weather	N.A.
McCabe and Trevathan, 2008	Premier League	Points for and against, win-loss record, home and away Performance, performance in previous 4 games, ranking, location, player availability	Richer feature sets	54%
Huang and Chang, 2010	World Cup 2006	Goals for, Shots, Shots on Goal, Corner Kicks, Free kicks, Ball possession & Fouls	Limited training data	77%
Tax and Joustra, 2015	Dutch Eredivisie	Goals for, goals against, result previous matched, top scorers, days since previous match, win/draw/lose percentage, Odds	Expand to other Leagues	55%
Aslan and Inceoglu, 2007	Italian Serie A	Home rating & Away rating for home team and away team	Different leagues, structures and input features	51%
Aslan and Inceoglu, 2007	Italian Serie A	Home Rating home team, Away Rating away team	Different leagues, structures and input features	53%

The McCabe and Trevathan paper from 2008, *"Artificial Intelligence in Sports Prediction"*, is considered a key milestone in sports prediction. It was able to achieve a 54% prediction accuracy for the Premier League (Table 1). A unique statistic the paper determined was the amount of "star" players available for a match, where a player is considered a "star" if they are currently involved in their nation's national team [6]. The 77% accuracy for the 2010 work by Huang and Chang in Table 1 cannot be considered with the other accuracies, because it used data from a World Cup, so some matches only had two outcomes (no draws). Another important note is that these past works, including Samba's, only used data from a single league.

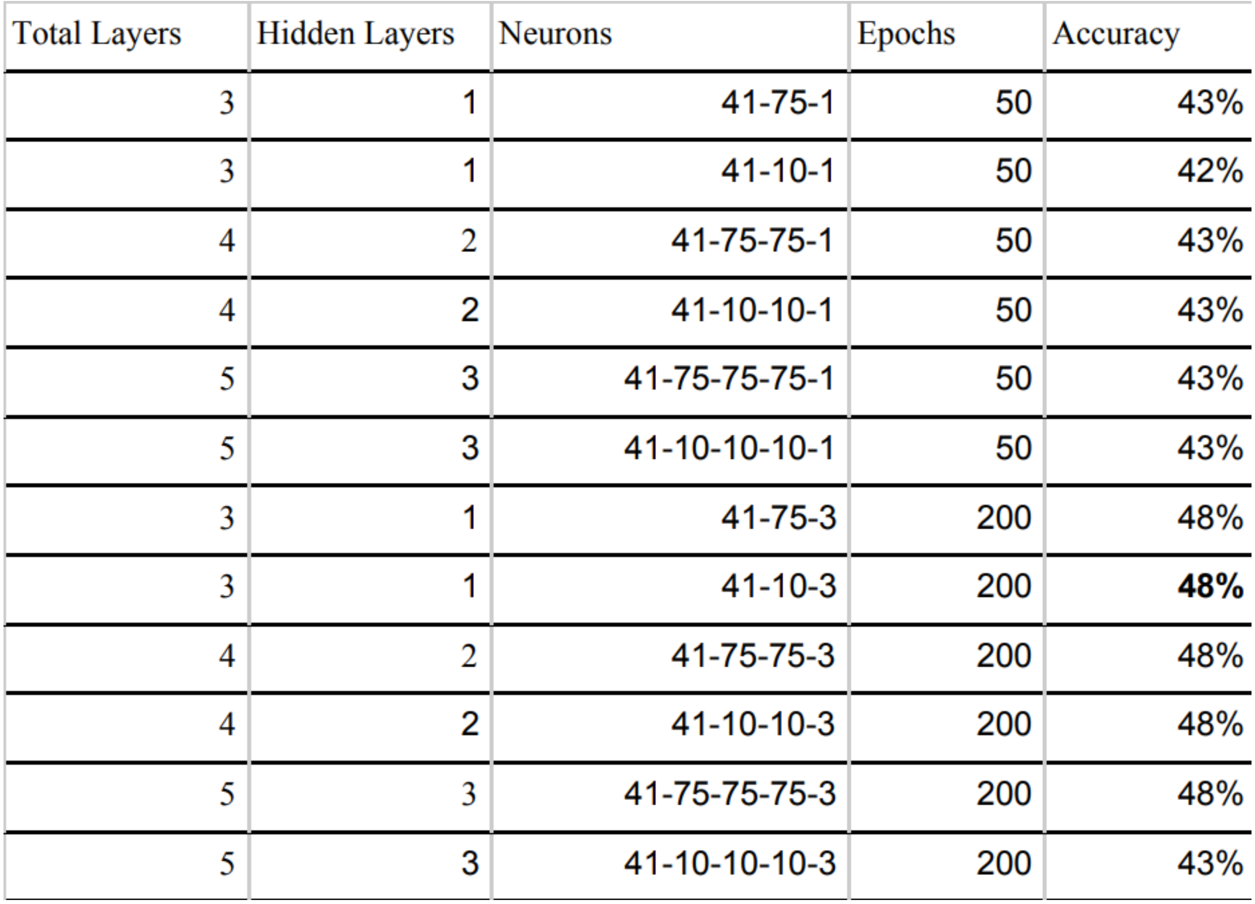
Table 2: An overview of network architectures reviewed by Samba [5]



Study	Total Layers	Hidden Layers	Neurons
Arabzad et al., 2014	4	2	10-20-20-2
McCabe and Trevathan, 2008	3	1	20-10-1
Huang and Chang, 2010	3	1	8-11-1
Tax and Joustra, 2015	N.A.	N.A.	N.A.
Aslan and Inceoglu, 2007	3	1	4-125-1
Aslan and Inceoglu, 2007	3	1	2-25-1

All of these networks use a Multi-Layer-Perceptron architecture. Looking at table 2, most of the past studies used a single-neuron output. However, Samba determined using three outputs is better (Table 3), meaning it is preferable to consider the three outcomes as separate probabilities.

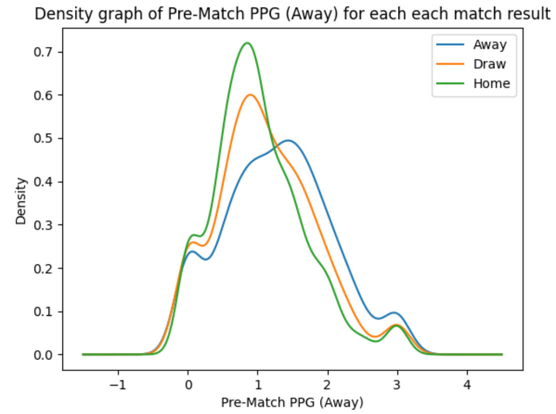
Table 3: Samba’s different network architectures and corresponding accuracy [5]



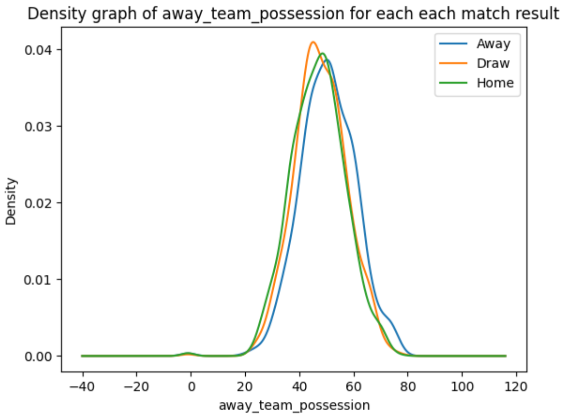
Total Layers	Hidden Layers	Neurons	Epochs	Accuracy
3	1	41-75-1	50	43%
3	1	41-10-1	50	42%
4	2	41-75-75-1	50	43%
4	2	41-10-10-1	50	43%
5	3	41-75-75-75-1	50	43%
5	3	41-10-10-10-1	50	43%
3	1	41-75-3	200	48%
3	1	41-10-3	200	**48%**
4	2	41-75-75-3	200	48%
4	2	41-10-10-3	200	48%
5	3	41-75-75-75-3	200	48%
5	3	41-10-10-10-3	200	43%

Data and Data Processing

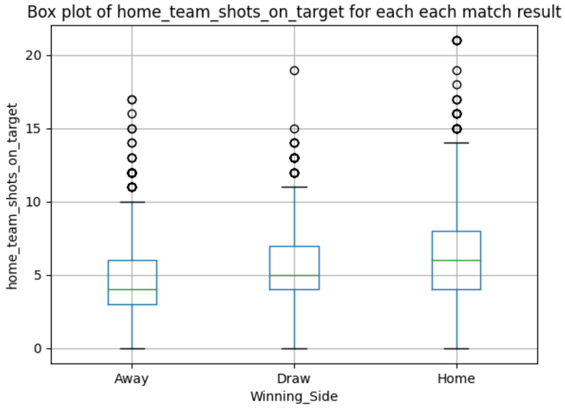
We downloaded CSV datasets from 5 different European soccer leagues, spanning the 2012-2020 seasons [7]. These CSV datasets consisted of 64 stats recorded for each match. We picked the 12 best stats using density plots and box plots which show the statistical relevance of a given input on match outcomes. We chose inputs that maximized the difference between the plots for each of the 3 match outcomes. These choices also agreed (mostly) with our intuition as soccer fans.



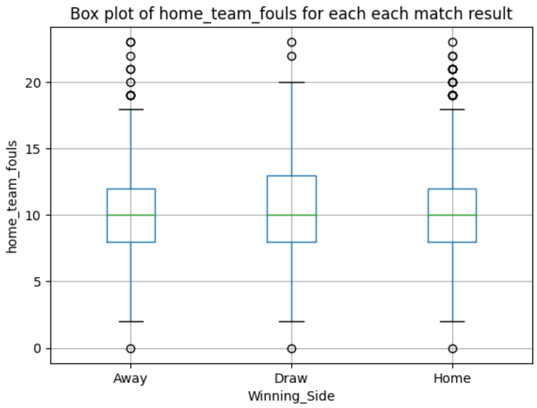
Density plot for a good input choice.



Density plot of a poor input choice.

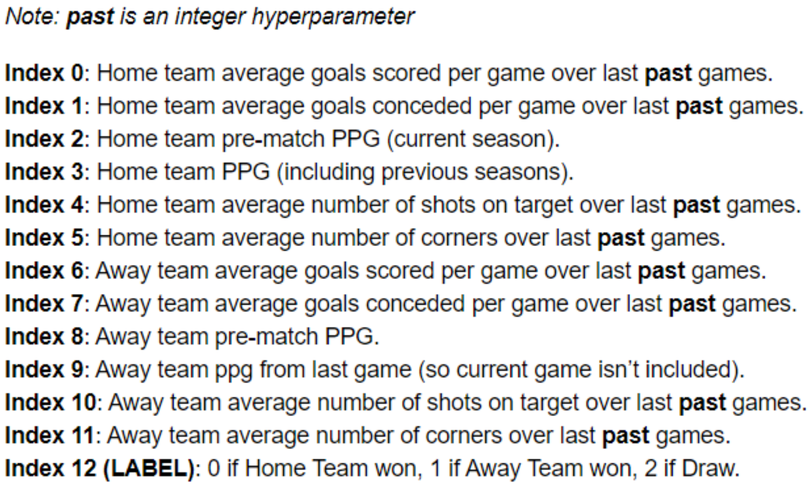


Box plot for good input choice.



Box plot for poor input choice.

So after choosing our 12 inputs we created a vector, as shown below, for each match in the 40 total seasons we used, then concatenated these vectors into one large dataset.

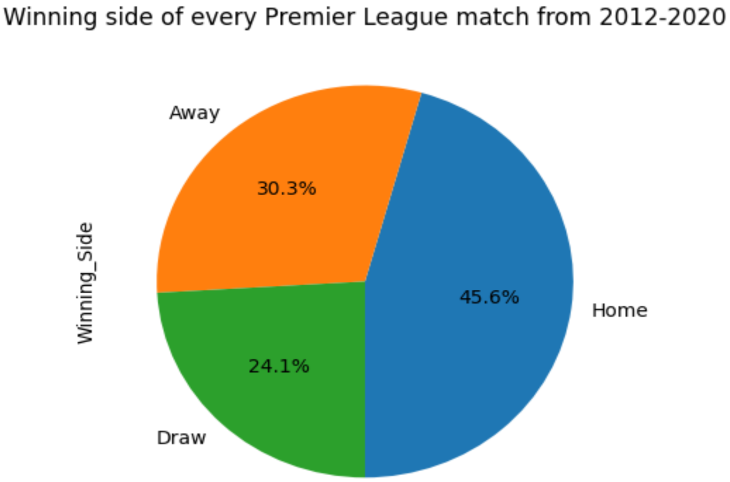


*Note: **past** is an integer hyperparameter*

Index 0: Home team average goals scored per game over last **past** games.
Index 1: Home team average goals conceded per game over last **past** games.
Index 2: Home team pre-match PPG (current season).
Index 3: Home team PPG (including previous seasons).
Index 4: Home team average number of shots on target over last **past** games.
Index 5: Home team average number of corners over last **past** games.
Index 6: Away team average goals scored per game over last **past** games.
Index 7: Away team average goals conceded per game over last **past** games.
Index 8: Away team pre-match PPG.
Index 9: Away team ppg from last game (so current game isn't included).
Index 10: Away team average number of shots on target over last **past** games.
Index 11: Away team average number of corners over last **past** games.
Index 12 (LABEL): 0 if Home Team won, 1 if Away Team won, 2 if Draw.

Finally we normalized each of the 12 vector entries (mean=0, std=1). While creating the above vector, we noticed that 3 of the leagues had missing 'Shots on Target' and 'Corners' data for many games in the 2012 season so we removed those games. Additionally, some matches were recorded out of chronological order so we had to sort our data by Match Week in order for us to be able to calculate average stats over a given number of *past* games.

Before class-balancing our dataset, the home team advantage phenomenon was clearly demonstrated:



This skewed dataset contained 12453 matches. After balancing, we had 9468 matches. We split this remaining data accordingly:

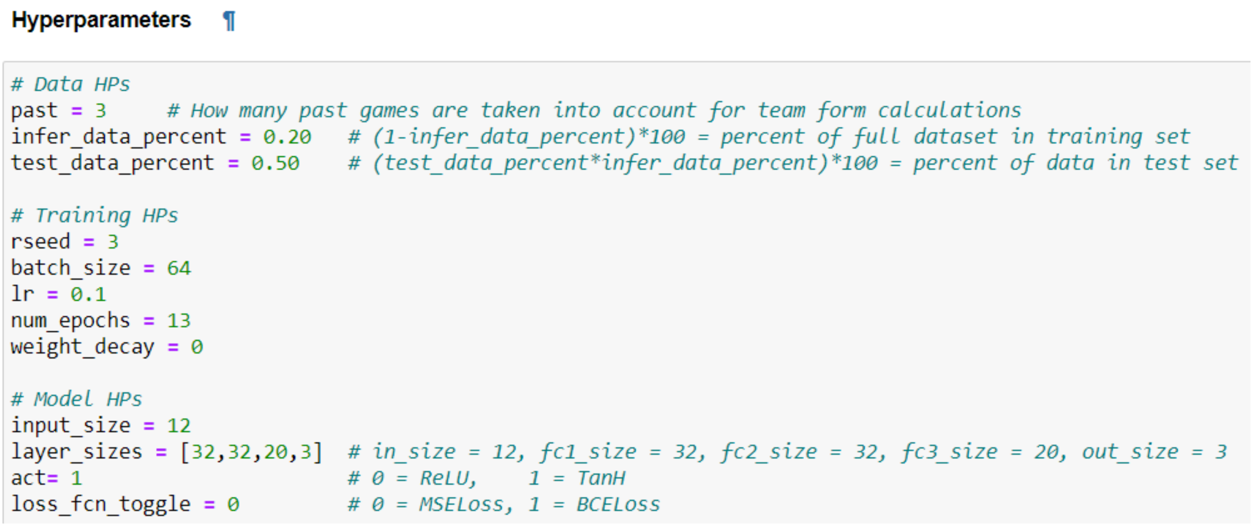
80% training set -- 10% validation set -- 10% test set
(7574 matches) (947 matches) (947 matches)

Architecture

Our model is an MLP which takes in an input vector of size 12 and outputs 3 numbers, each representing the probability (or more precisely, the model's confidence) of a given outcome occurring. This probability-like output is created using the SoftMax activation function on the output layer. We used Stochastic Gradient Descent as our optimizer because it performs well and we are comfortable using it. The rest of the hyperparameters, which we optimized using a manual grid search are shown below.

Model With Best Performance

→ No weight decay, early stopping (after 13 epochs).



Hyperparameters ¶

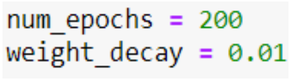
```
# Data HPs
past = 3     # How many past games are taken into account for team form calculations
infer_data_percent = 0.20    # (1-infer_data_percent)*100 = percent of full dataset in training set
test_data_percent = 0.50     # (test_data_percent*infer_data_percent)*100 = percent of data in test set

# Training HPs
rseed = 3
batch_size = 64
lr = 0.1
num_epochs = 13
weight_decay = 0

# Model HPs
input_size = 12
layer_sizes = [32,32,20,3]  # in_size = 12, fc1_size = 32, fc2_size = 32, fc3_size = 20, out_size = 3
act= 1                      # 0 = ReLU,     1 = TanH
loss_fcn_toggle = 0         # 0 = MSELoss, 1 = BCELoss
```

Model Using Weight Decay With Best Performance

→ All hyperparameters the same as above, except for...



```
num_epochs = 200
weight_decay = 0.01
```

We settled on 4 fully connected layers with the given sizes because:

1. This model produced equal or better results to any other architecture we tried.
2. The successful reference models we looked at had 3-5 layers with 3 output neurons on the last layer.
3. Rule of thumb: "Number of trainable parameters should be *at most* equal to the training set size for successful training to be likely."
   We carried out our hyperparameter search using a reduced training set containing 2200 examples (8 seasons worth of matches), so we specified the number of parameters to be below this. Layer sizes of [32,32,20,3] plus 12 inputs produces a model with 2195 total parameters, so it satisfies the heuristic.

Baseline Model

- The simplest model which addresses this problem randomly guesses between the three outcomes. This yields an accuracy of 33% on a balanced dataset.
- A better model which always picks “Home Team Win” achieves an accuracy of 45.5% on our dataset before class-balancing. This accuracy reflects a more realistic distribution of results, where home team advantage is a proven phenomenon. Even though our model was trained and tested on a balanced dataset, if the test accuracy is not above this benchmark there is no value in using our neural network.

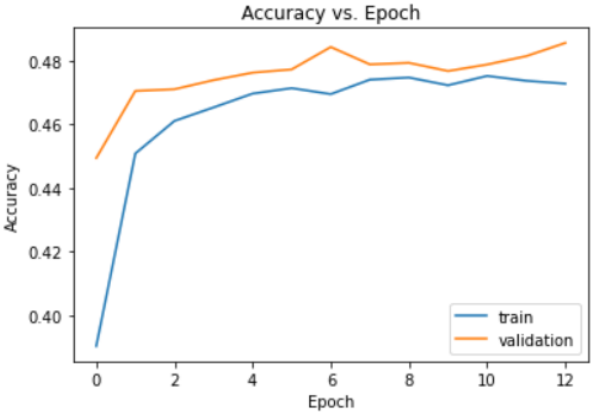
Quantitative Results

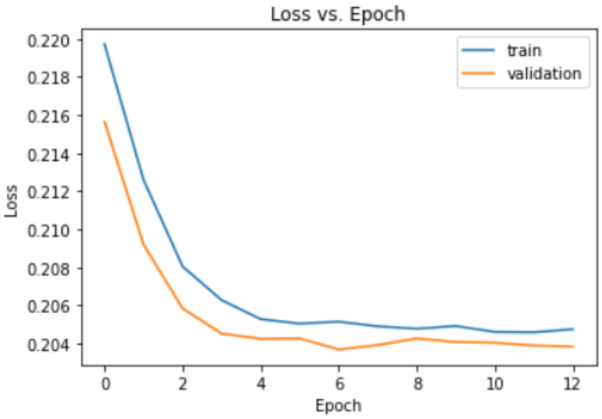
Within the given context of sports betting, the only metrics that matter are validation and test accuracies. The intended use of our model is to make a prediction that a sports fan could use for betting. If our model is incorrect, the type of incorrect does not concern them because the outcome is still the same, they lose money. The other models also use accuracy as the sole performance metric [5].

Model with Best Performance

Max. Training Accuracy: **47.5%**	Min. Training Loss: **0.2046**
Max. Validation Accuracy: **48.6%**	Min. Validation Loss: **0.2037**
Max. Test Accuracy: **47.5%**	Min. Test Loss: **0.2043**
Total training time: **11.7 seconds**	

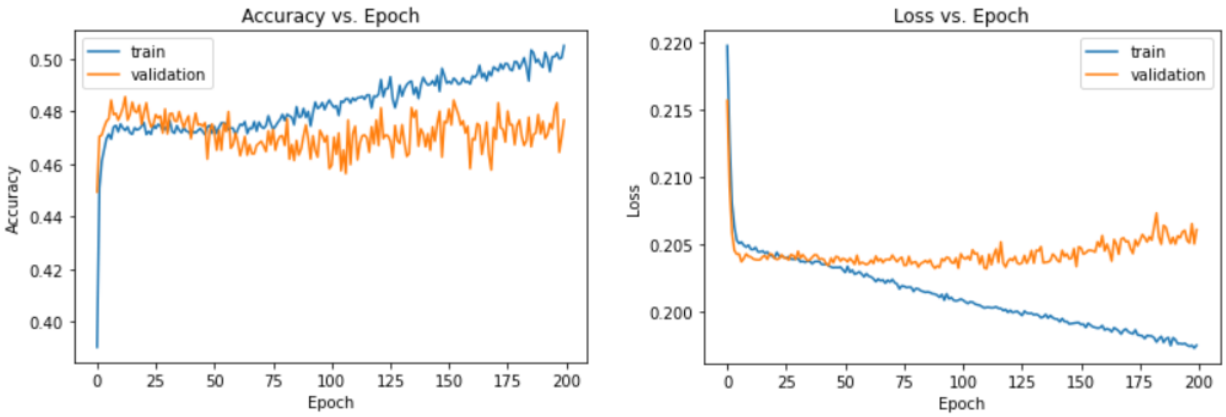
Baseline model beaten!





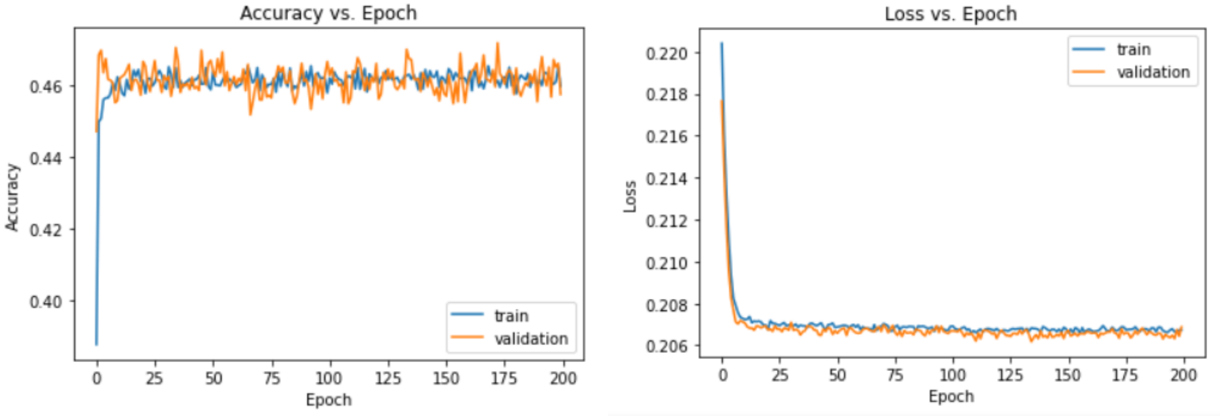
Overfitting Issues

→ In later epochs, interesting training breakthroughs could happen.
→ Overfitting causes validation accuracy to drop and training longer is useless.



Weight Decay

→ Implemented weight decay to combat overfitting.

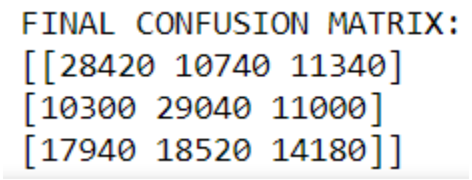


Overfitting problem solved! *But overall performance suffers*.

Max. Validation Accuracy: **47.2%** Max. Test Accuracy: **46.7%**

Therefore we went with no weight decay and early stopping to achieve our best results.

Qualitative Results



```
FINAL CONFUSION MATRIX:
[[28420 10740 11340]
[10300 29040 11000]
[17940 18520 14180]]
```

Note 1: Entry $C_{i,j}$ is the number of examples known to be in class *i* and predicted to be in class *j*.

Note 2: Classes are Home Team Win (top/left), Away Team Win (middle), Draw (bottom/right).

Our model performs poorly when predicting draws, with away win predictions effectively as accurate as home win predictions. The difficulty with predicting ties is there are multiple scenarios that create a tie; evenly matched teams or a significant drop in performance from the better team according to bookie strategies [8]. The two scenarios are dependent on the competitiveness of the league. Our hypothesis is that the model learned one strategy, but had trouble with the variation in the different leagues. A second factor is the motivation for both teams to get a result. Towards the end of a season, teams can be satisfied with a draw given their position in the league is safe. Our model has no input corresponding to the gameweek, therefore there is no “motivation” statistic.

Discussion and Learnings

While our final test accuracy (47.2%) is just below the final accuracy Samba was able to produce in their neural network (48%), ours is generalized to work with the five most popular leagues in Europe, while Samba’s was trained only using the English Premier League [5]. In fact, all the other neural networks used in past studies were trained using a single league’s data (Table 1). It is harder to create a model which works for more than one league because different leagues have certain “play styles” unique to them. For example, teams in the Italian league, *Serie A*, have historically been known to be very defensive, and as a result the average goals per game in a *Serie A* match is lower than other leagues [9]. Because of this, we believe our model performed well, since it almost matched the accuracy of a model which used only a single league’s data. Our group saw value in attempting to create a more generalized model, since it’s been rarely attempted and would have more potential uses.

Figure 1: Heat map of soccer match

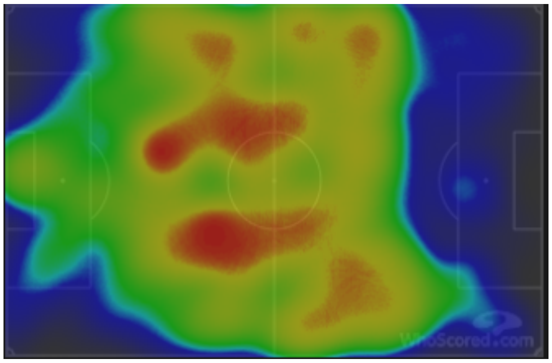
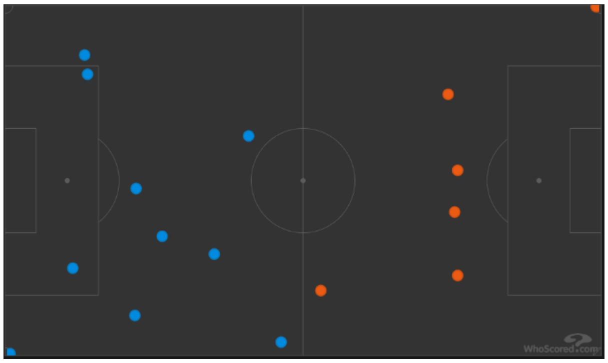


Figure 2: Location of key passes



If we were to start another similar project from scratch, we would include more detailed match data, such as heat maps (spatial histograms as depicted in Figure 1), key pass locations (Figure 2), player lineups and individual statistics. Team statistics can often be broken down per player, which would be useful when players miss games or move to a different team mid-way through a season. Mccabe and Trevathan were able to reach 54% accuracy with their model back in 2008 which included such statistics [6]. To account for more input features, we would need more sample matches to avoid over-fitting, but since we're creating a generalized predictor, we can simply obtain more data from the thousands of other soccer leagues around the world.

Ethical Framework

The most obvious stakeholders for a project like this would be soccer fans and sports betting companies. Fans could use this tool recreationally to get a better understanding of the relative strength of teams in their favourite league, while betting companies could potentially make or lose a large amount of money, depending on whether they or their customers have access to a more accurate match predictor. Some less obvious stakeholders include soccer analysts, whose jobs could be at risk if a neural network is able to provide more accurate predictions than they can, and the soccer teams themselves (players, coaches, staff), since the more accurate a predictor such as this is, the greater potential for them to rely on it for making decisions.

We will now look at how Reflexive Principlism applies to soccer fans and sports betting companies.

Soccer Fans:

Non-maleficence: Although not obvious, careful thought must be put into the way data & predictions are presented to fans to avoid causing harm. Rowdy soccer fans, known as “hooligans”, have always been a problem in European leagues, especially in Russia and the United Kingdom. Modern day “hooliganism” has taken the form of verbal and online harassment, often of racist nature [10]. A disapproved team or player by the model that also happens to represent a minority group could be used to fuel more racist remarks by hooligans.

Beneficence: The main purpose of this tool would be to provide benefits to the fans in the form of knowledge and insight into how their favourite teams will perform. The more transparent the model is, the more information fans can take away.

Betting Companies:

Justice: With the $250 billion sports betting industry being made up of many powerful companies [3], it is important to consider fairly (or unfairly) distributing the benefits and risks of a tool such as this. A more accurate predictor could allow a company to rise above the rest if they are given exclusive rights, so careful thought should be on whether to make this tool open-source.

Non-maleficence: While this tool has the potential to earn betting companies more money, it can also cause harm by losing them more money with inaccurate predictions. This could result in many lost jobs and negatively affect their stock price. Because of the massive power and wealth of some of these companies, it opens the possibility of lawsuits against the creators of the model.

References

[1] J. Shvili, “The Most Popular Sports In The World,” WorldAtlas, 16-Oct-2020. [Online]. Available: https://www.worldatlas.com/articles/what-are-the-most-popular-sports-in-the-world.html. [Accessed: 04-Dec-2020].

[2] C. Gough, “Sports broadcasting rights by league 2019/20,” Statista, 02-Jun-2020. [Online]. Available: https://www.statista.com/statistics/1120170/broadcasting-rights-sports-by-league/. [Accessed: 04-Dec-2020].

[3] A. Gray, “The Size and Increase of the Global Sports Betting Market,” *Sports Betting Dime*, Apr-2018. [Online]. Available: https://www.sportsbettingdime.com/guides/finance/global-sports-betting-market/. [Accessed: 04-Dec-2020].

[4] Yadav A, Sharma A, Gautam A, Bathla G, Jindal R (2017) Predicting English Premier League Results using Machine Learning. J Comput Eng Inf Technol 6:1. doi: 10.4172/2324-9307.1000165

[5] Samba, Stefan. (2019). Football Result Prediction by Deep Learning Algorithms. 10.13140/RG.2.2.25014.45122.

[6] Mccabe, Alan & Trevathan, Jarrod. (2008). Artificial Intelligence in Sports Prediction. 1194-1197. 10.1109/ITNG.2008.203.

[7] “Download Soccer / Football Stats Database to CSV: FootyStats,” *Football Stats by FootyStats*. [Online]. Available: https://footystats.org/download-stats-csv. [Accessed: 25-Nov-2020].

[8] Bobslay, “Using Historic Odds Trends To Predict Draws - bobslay13 Blog,” OLBG.com - Let's Beat The Bookies, 2016. [Online]. Available: https://www.olbg.com/blogs/using-historic-odds-trends-predict-draws. [Accessed: 05-Dec-2020].

[9] B. McAleer, “League Focus: Is Serie A Still The Most Defensive League In Europe?,” es.whoscored.com, 29-Apr-2013. [Online]. Available: https://es.whoscored.com/Articles/M9aoDg2rEkCE032scMMbyw/Show/League-Focus-Is-Serie-A-Still-The-Most-Defensive-League-In-Europe. [Accessed: 05-Dec-2020].

[10] S. Parkin, “The rise of Russia's neo-Nazi football hooligans,” The Guardian, 24-Apr-2018. [Online]. Available:

https://www.theguardian.com/news/2018/apr/24/russia-neo-nazi-football-hooligans-world-cup. [Accessed: 05-Dec-2020].

Permissions

Name	Permission to post video	Permission to post final report	Permission to post source code	Initials
Atom Arce	Yes	Yes	Yes	AA
Spencer Ball	Yes	Yes	Yes	SB
Markus Kunej	Yes	Yes	Yes	MK