1. Introduction

Predicting the outcome of sports matches is one of the favorite topics for statisticians and gamblers alike. It is an interesting problem where the question of the respective importance of randomness ('luck') and determinism ('skill') is hotly debated. Of course, much importance also lies in the fact that successful predictions can lead to lots of money. For soccer, previous prediction methods mostly used team-level features while neglecting individual skills. We hypothesize that individual skills are crucial to soccer and contain much information for predicting match outcomes which those other methods are not utilizing. Therefore, in this project, we sought to develop a machine learning strategy that predicts the outcomes of soccer matches based mainly on the individual attributes of the players on each team.

2. Data Collection and Processing

We have implemented a vast array of programs that facilitate the process of scraping data from different websites and then pre-processing and combining them to yield consistent datasets that can be used for machine learning.

Our novel approach uses as features the in-game stats from the Electronic Arts' celebrated game franchise FIFA, which were painstakingly crafted by the experts. Those numerical stats (1-100) measuring skills falling under broad categories (attacking, skill, movement, power, mentality, defending, and goalkeeping) were obtained from ¡https://sofifa.com¿. For practical reasons, we averaged over different positions in the team (forward, midfielder, defender, goalkeeper) all the skills falling under those categories. This was then combined with the records of matches (obtained from ¡http://www.worldfootball.net¿) to yield a dataset consisting of match outcomes and player stats. For this project, we focused on Premier League matches played in 2010-2017. After throwing away data entries plagued with various errors and difficulties, we ended up with a training set consisting of 2104 Premier League matches played in 2010-2017 minus the 2015-2016 season and a test set consisting of 273 matches played in the 2015-2016 season.

This was all done by writing a vast array of programs that facilitate all the steps from scraping raw xml off of different websites to processing and combining them to yield consistent datasets that can be used for machine learning.

3. Data Analysis Approach

The first method that we used to analyze the dataset obtained by above procedure was Weka. We used various algorithms that were identified to be promising during our preliminary investigations. Specifically, those included three tree algorithms, three Bayesian algorithms, and two logistic regression algorithms.

Complementing the Weka, we decided to create our own machine learning implementation. Specifically, we wrote a Python script implementing a neural network with one hidden layer and three softmax output units each corresponding to the probability of the input data belonging to each class. The algorithm also uses 10-fold cross-validation with random partitioning of the dataset into training/validation sets, intentionally mirroring the Weka's approach. One thing of note is that we have normalized the data such that each attribute lies in the $[0, 1]$ range (i.e. divided by 100). This seemed to improve the result significantly (increasing the classification accuray by 5-10%), presumably due to the reduction of overflow errors. For Weka performance, it had no significant impact.

Finally, we sought to investigate the relative predictive power of each feature. This is not only interesting in its own right (we might be able to determine which skills are most relevant for soccer), it can also be a helpful aid for guiding machine learning implementation, in particular serving as a sanity-check for the approach. In order to measure this quantitatively, we wrote a Python script that calculates, for each attribute, the maximum information gain that can be achieved by making a split along that attribute (cf. decision trees).

4. Results

As mentioned above, we used 2104 Premier League matches played in 2010-2017, minus the 2015-2016 season, as our training and cross-validation set. 273 matches played in the 2015-2016 season was set aside as the test set.

Table 1 shows the results of various Weka algorithms. Sadly, none of the algorithms performed significantly better than the ZeroR baseline, measured in terms of their 10-fold cross-validation accuracy. Logistic regression, which performed the best, only yielded a 3

Algorithm	Training accuracy	10-fold cross-validation accuracy	Test accuracy
ZeroR	0.4644	0.4644	0.3919
J48	0.9439	0.3912	0.3993
REPTree	0.5741	0.4838	0.4505
RandomForest	1.0000	0.4829	0.4396
NaiveBayes	0.4049	0.3593	0.3773
BayesNet	0.4967	0.4829	0.4432
NaiveBayesMultinomial	0.4316	0.4335	0.4029
Logistic	0.5318	0.4948	0.4286
SimpleLogistic	0.5119	0.4924	0.4139

Table 1: Training and 10-fold cross-validation accuracies for Premier League matches in 2010-2017. 2015-2016 season data was set aside as the test set to derive the test accuracy. The highest accuracy in each column has been shaded in light red.

Our custom single hidden layer neural network implementation was no more successful. Figure 4.1 shows the learning curve of the network with 30 hidden units. It is clear that there is underfitting and the model is incapable of capturing the underlying trend of the data, if there even is any. Same results were obtained for hidden layer size ranging from 10-100 units (see Figure 4.2). This suggests that either the neural network architecture considered is insufficient (for instance, more than one hidden layers may be necessary) or, as is more likely in the light of the failures of other algorithms above, the problem lies with the dataset itself.

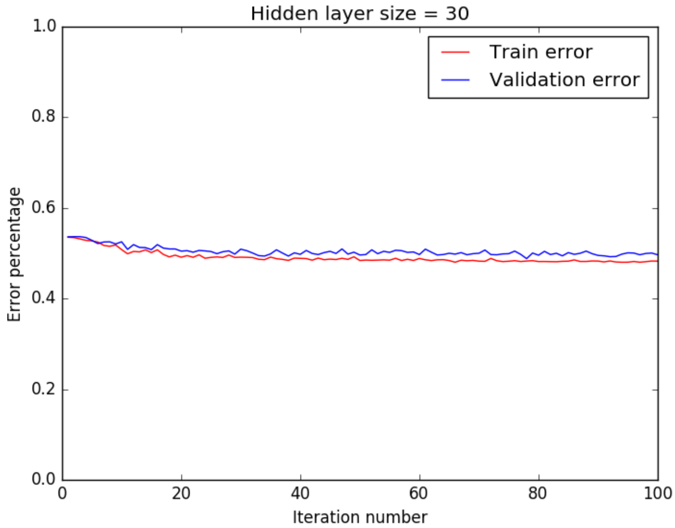

Figure 4.1

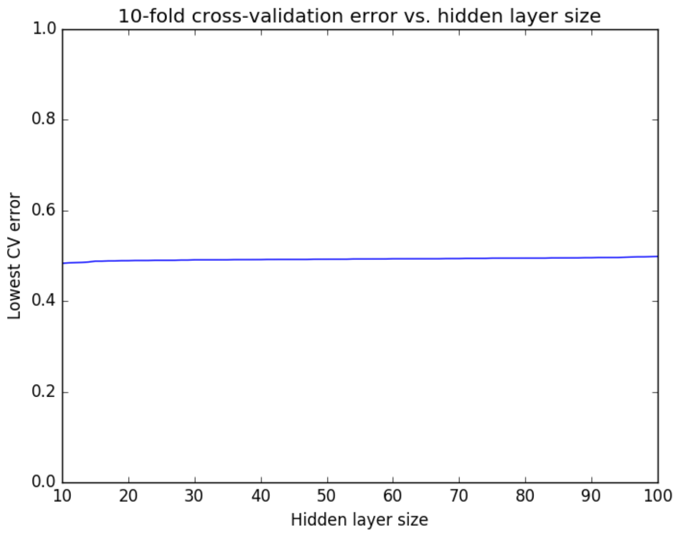

Figure 4.2

Finally, Table 2 lists select features and the corresponding information gain that can be obtained by making a split over the feature. Full data is available in the repository. A cursory glance suggests that skills related to goalkeeping and defending are the most important factors in deciding the outcome of a soccer match. However, this result should be taken with a grain of salt as there are also nonsensical results such as the goalkeeper's attacking or the forward's goalkeeping skills (which should be completely irrelevant to the match) yielding more information than defender's defending skills (which is clearly relevant). This further suggests that our data is seriously flawed.

No.	Attribute	Information gain
1	Home Team; Goalkeeper; Goalkeeping	0.03567655
2	Away Team; Goalkeeper; Goalkeeping	0.024776643
3	Home Team; Defender; Defending	0.013332217
4	Home Team; Goalkeeper; Movement	0.013321315
5	Home Team; Defender; Movement	0.011140302
⋮	⋮	⋮
37	Home Team; Goalkeeper; Attacking	0.003304952
⋮	⋮	⋮
43	Away Team; Goalkeeper; Attacking	0.002832013
⋮	⋮	⋮
51	Home Team; Forward; Goalkeeping	0.002177446
52	Away Team; Forward; Goalkeeping	0.002115403
53	Away Team; Defender; Defending	0.00206549

Table 2: A selection of attributes listed in the order of decreasing information gain (in bits). Prior entropy is 1.5305276569.

5. Discussion

In retrospect, we may have been overly optimistic and ambitious in pursuing this project. Of course, soccer is a fundamentally team-based sport and neglecting crucial factors such as formations and previous team records may have doomed the prospect of success from the beginning. Furthermore, quantitative measures of individual attributes, including video game stats used here, must by necessity incorporate some subjective judgments and may even be inconsistent from year to year.

There are of course various alternative ways of formulating features from individual numerical stats that can be tried. However, considering that our current approach has yielded such abysmal results, we do not expect to obtain drastically improved results following this approach. It is likely that our data is fundamentally limited and it is probably better to find a different data source altogether for measuring individual skills.

Finally, it may be the case that although individual attributes alone do not have enough predictive power, combining them with team-based features can yield drastically improved performance that cannot be achieved by using either set of features alone. This would be the natural direction to explore in any future work.

6. Conclusion

In conclusion, Soccer Success.

7. Acknowledgements

Jeremy Rath worked mainly on data collection and processing. Hyun Jin Kim worked mainly on data analysis. Both authors worked on setting up the webpage and preparing this report.