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

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

Position-averaged, categories.

All Premier League seasons sans 2015-2016 (set aside as the test set).

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

Three differeng leagues.

Algorithm	Premier League	Bundesliga	La Liga
ZeroR	0.4801	0.4896	0.4878
J48	0.4709	0.3889	0.4177
REPTree	0.4618	0.4514	0.5213
RandomForest	0.5505	0.4896	0.4787
NaiveBayes	0.3700	0.3299	0.2927
BayesNet	0.4954	0.4896	0.4360
NaiveBayesMultinomial	0.4037	0.3125	0.4024
Logistic	0.4771	0.4479	0.4390
SimpleLogistic	0.5260	0.5000	0.5183

Table 2: 10-fold cross-validation accuracy for 2016-2017 season for three different leagues. The highest accuracy for each league has been shaded in light red.

Neural network

We have trained models using different hidden layer sizes ranging from $10 - 100$. One example of a learning curve (for hidden layer size $= 30$) is shown in Figure 4.1. It can be seen clearly that the model is incapable of fitting the dataset. In fact, essentially the same results were obtained for all values of hidden layer size used (see Figure 4.2). While this might just suggest that the neural network architecture considered is inappropriate for this problem (for instance, more than one hidden layers may be necessary), considering that none of the other algorithms used above in Weka yielded much success, it seems more likely that the problem lies with the dataset itself.

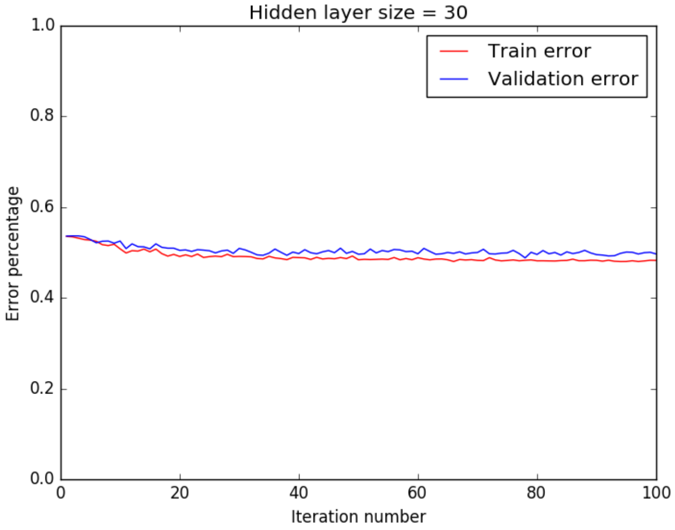

Figure 4.1

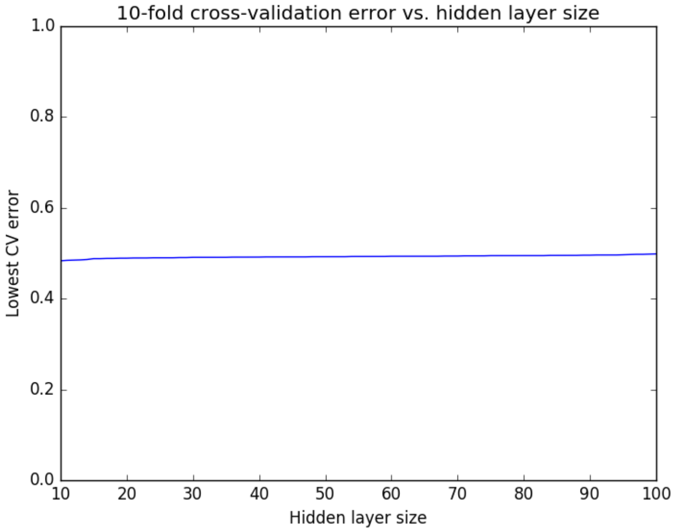

Figure 4.2

Feature analysis

Table 3 lists some of the features with the corresponding information gain, with more informative features towards the top. Full data is available in the repository.

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 3: Prior entropy = 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.

6. Conclusion

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.