Gaining a Statistical Edge in Soccer Prediction using Machine Learning: Role of Meta Statistics in Match Prediction

Varun Harbola[1] and Kyuho Lee[1]

1. Department of Physics, Stanford University, Stanford, CA 94305

1. Background

Sports match prediction is a huge industry, spanning from pre-/post-match analysis to sports betting and sports management.[1] In predicting the outcome of a match between two teams, statistics such as historical match results between the two teams have been conventionally employed. However, a match outcome is strongly dependent on a much wider spectrum of team statistics and tactics employed during the match. To navigate in this complex feature space, we employed machine learning algorithms on various sets of relevant features for English Premier League (EPL) soccer match prediction.

- English Premier League soccer match prediction
- Blind Guess Accuracy: 33.33% (win/draw/loss)
- Home-Team-Win Guess Accuracy: 46.19%
- Literature Accuracy (Neural Network): 53.25%[2]
- **Goal:** surpass these benchmark values.

2. Dataset and Feature Selection

- data scraping from https://www.whoscored.com
- 2009/2010 ~ 2018/2019 EPL season matches: 3800 matches in total
- 3 different sets of features:
  - home & away average roster rating (F1)
  - home & away average roster rating + positional average rating (F2)
  - individual player performance metrics for all players in home & away team (F3)
    (overall rating, pass success percentage, passes per game, shots per game, key passes per game, blocks per game, interceptions per game)
- 80% and 20% of randomly sorted data allocated to training & test data sets
  (features are time-independent, so we assume model is also time-independent)

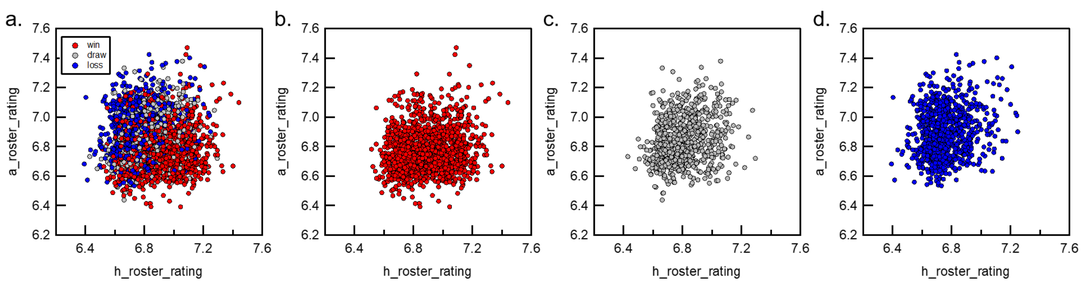


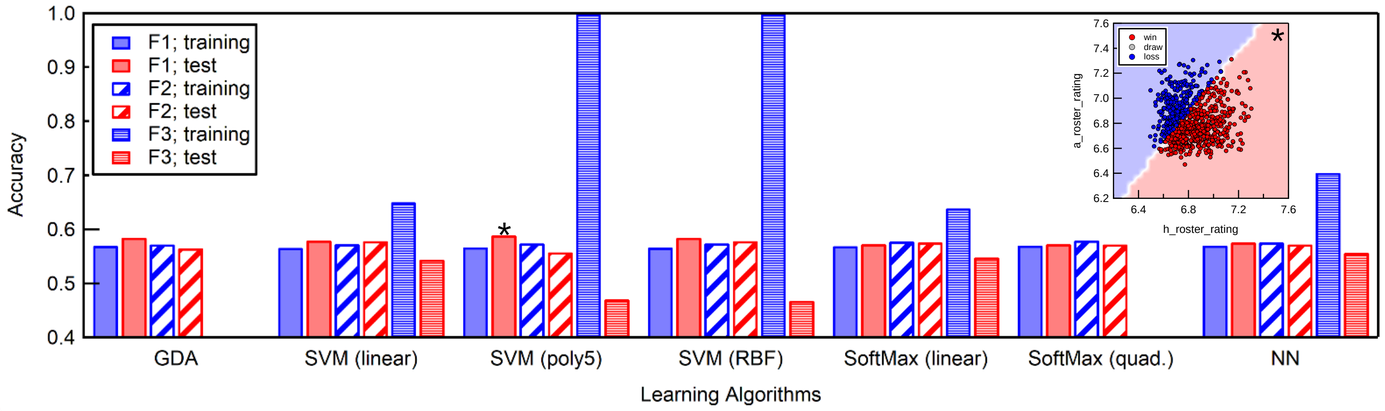
Figure 1. a) Plot of the training data with the home and away average roster ratings on the *x*-axis and the *y*-axis, respectively, along with individual plots of **b**) home win, **c**) draw, and **d**) home loss datapoints.

3. Learning Algorithms & Results

Models Used:

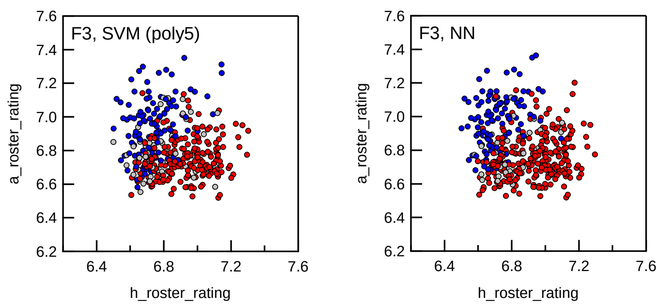
- Gaussian Discriminative Algorithm (GDA)
- Support Vector Machine (SVM) with linear, 5-degree polynomial, and radial basis function kernels
- SoftMax Regression with linear and quadratic features
- Neural Network (NN) with 1 hidden layer and 11 hidden nodes

58.89% accuracy achieved for degree-5 polynomial kernel SVM with feature set F1.



4. Limitations in Draw Game Predictions

Improvement in draw prediction is achieved only by overfitting on the draw games in F3. This causes the overall accuracy to drop on test set predictions, even though draws are now being accounted for in the model.



5. Conclusion

We achieved prediction accuracy of up to 58.89%, surpassing the literature values by ~10%.[2] Draw prediction is improved by enlarging feature space, but overfitting problems occur simultaneously. These results call for further developments in learning algorithms to improve draw game predictions.

6. Reference

1 Lopez-Gonzalez, H. & Griffiths, M. D. Understanding the Convergence of Markets in Online Sports Betting. *International Review for the Sociology of Sport* **53**, 807 (2018).

2 Aslan, B. G. & Inceoglu, M. M. in *Seventh International Conference on Intelligent Systems Design and Applications (ISDA 2007)*. 545.