SJIF Impact Factor: 7.001| ISI I.F.Value:1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 5 | Issue: 10 | October 2020 - Peer Reviewed Journal

TRENDS OF DEVELOPMENT IN MACHINE LEARNING TECHNIQUES FOR SOCCER PREDICTION

Ajayi Oluwabukola

Ajayi Adebowale

ABSTRACT

With the vast amounts of available data in the football industry researchers have applied several machine learning techniques towards predicting soccer outcomes. Feature engineering has been a major focus of these researches and yet the jury remains out for the best combination of data features and machine learning algorithms to be used for accurate predictive models.

This study presents a review of studies carried out in the intersection of machine learning and soccer prediction with a view to characterizing and presenting a trend of developments in this budding research area.

KEYWORDS: *football industry, machine learning, algorithms, feature engineering, soccer prediction*

INTRODUCTION

Data Science is presently on the forefront of the football industry with many possible uses and applications: Match strategy, tactics, and analysis, Identifying players' playing styles, Player acquisition, player valuation, and team spending, Training regimens and focus, Injury prediction and prevention using test results and workloads, Performance management and prediction, Match outcome and league table prediction, Tournament design and scheduling and Betting odds calculation. The focus of this research however is on the application of machine learning algorithms for the prediction of soccer matches.

Several researchers have proposed techniques for crunching soccer numbers in an attempt to predict games. The focus has been on identifying relevant factors to include in the predictive model, however results have been inconclusive and the jury is still out for improved methodologies for soccer prediction.

This study developed a soccer prediction framework that included individual player features and aggregated team features and implemented the developed framework by using an ensemble of support vector machines and logistic regression models on historical football data.

FOOTBALL PREDICTION LANDSCAPE

Generating predictions for football scores has been an important research theme since the middle of the 20th century, with the first statistical modelling approaches and insights coming from Moroney (1956) and Reep (1971) who used both the Poisson distribution and negative binomial distribution to model the amount of goals scored in a football match, based on past team results. However, it was only in 1974 that Hill proved that match results are not solely based on chance, but can be modeled and predicted using past data .

The first breakthrough came from Maher in 1982 who used Poisson distributions to model home and away team attacking and defensive capabilities, and used this to predict the mean number of goals for each team. Following this, Dixon and Coles (1997) were the first to create a model capable of outputting probabilities for match results and scores, again following a Poisson distribution.

The Dixon and Coles model is still seen as a traditionally successful model, and we will use it as a benchmark against the models that we will be creating. The Dixon and Coles model is based on a Poisson regression model, which means that an expected number of goals for each team are transformed into goal probabilities following the Poisson distribution (illustrated in Fig 1.0):

SJIF Impact Factor: 7.001| ISI I.F.Value:1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 5 | Issue: 10 | October 2020 - Peer Reviewed Journal

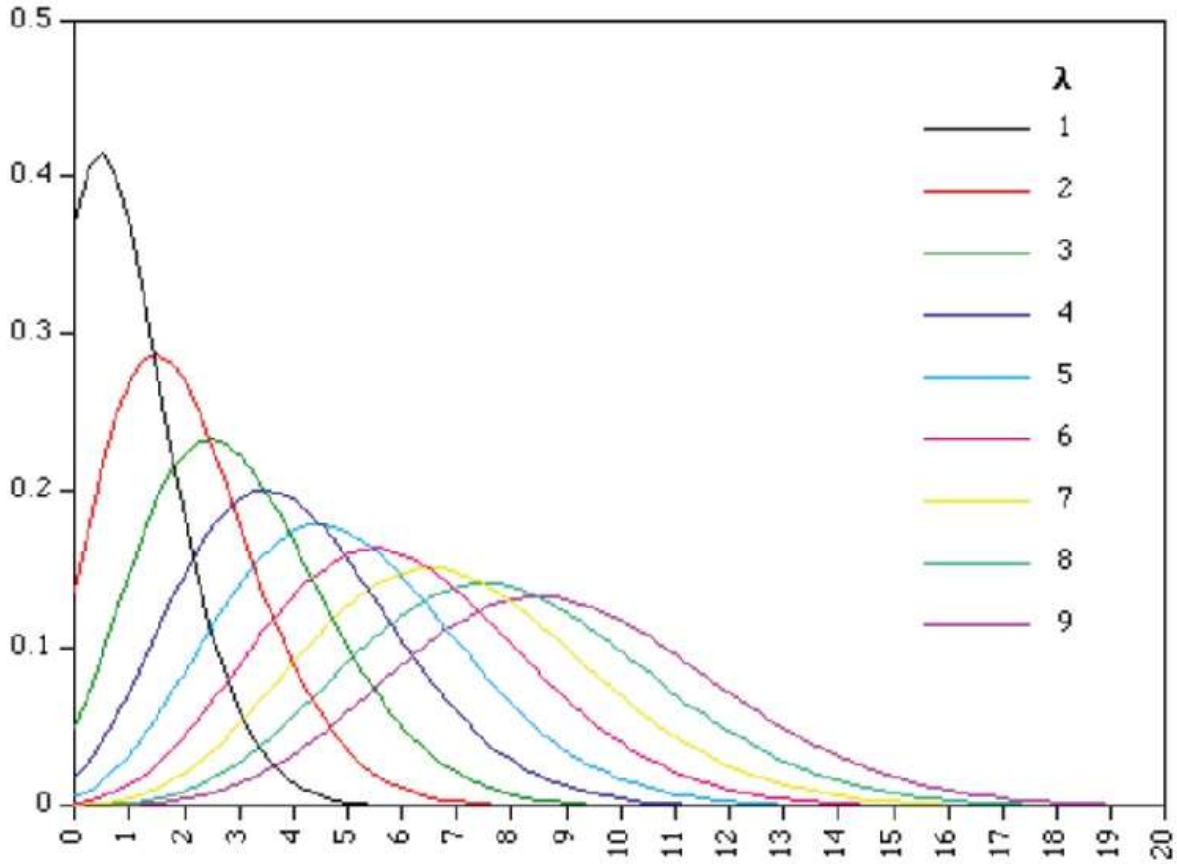

Figure 1: Poisson distribution for values of λ and k

The Dixon and Coles model is based on a Poisson regression model, which means that an expected number of goals for each team are transformed into goal probabilities following the Poisson distribution (illustrated in Fig.2.9):
P(k goals in match)= e−% %k k!
where % represents the expected number of goals in the match.

The Poisson distribution enables the calculation of the probability of scoring a certain number of goals for each team, which can then be converted into score probabilities and finally into match outcome probabilities.

Based on past results, the Dixon and Coles model calculates an attacking and defensive rating for each team by computing Maximum Likelihood estimates of these ratings on past match results and uses a weighting function to exponentially downweight past results based on the length of time that separates a result from the actual prediction time.

Rue and Salveson (2000) chose to make the offensive and defensive parameters vary over time as more results happen, then using Monte Carlo simulation to generate predictions. In 2002, Crowder et al. followed up on their work to create a more efficient update algorithm.

At the beginning of the 21st century, researchers started to model match results (win/draw/loss) directly rather than predicting the match scores and using them to create match result probabilities. For example, Forrest and Simmons (2000) used a classifier model to directly predict the match result instead of predicting the goals scored by each time. This allowed them to avoid the challenge of interdependence between the two teams' scores.

During the same year, Kuypers used variables pulled from a season's match results to generate a model capable of predicting future match results. He was also one of the first to look at the betting market and who tried to generate profitable betting strategies following the model he developed. We have therefore seen that past research have tried to predict actual match scores as well as match results. It would be interesting in this project to look at the performance of generating a classification model for match outcome against a regression model for match scores.

We will now take a look at more recent research done on the subject, with the use of modern Machine Learning algorithms that will be interesting for us to investigate when trying different predictive models.

In 2005, Goddard tried to predict football match results using an ordered probit regression model, using 15 years of results data as well as a few other explanatory variables such as the match significance as well as the geographical distance

SJIF Impact Factor: 7.001| ISI I.F.Value:1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 5 | Issue: 10 | October 2020 - Peer Reviewed Journal

between the two teams. It was one of the first papers to look at other variables than actual match results. He compared the model predictions with the betting odds for the matches and found that there was the possibility of a positive betting return over time. Like Goddard, we will want to use other explanatory variables in our model than only match results, which will allow us to use different sets of features to try obtaining the best model possible.

It is also interesting to look at the algorithms used for predictions in other team sports: for example, in 2006, Hamadani compared Logistic Regression and SVM with different kernels when predicting the result of NFL matches (American Football). More recently, Adam (2016) used a simple Generalised Linear Model, trained using gradient descent, to obtain match predictions and simulate the outcome of a tournament. He obtained good results, even with a limited set of features, and recommends to add more features and to use a feature selection process, which is something that will be interesting for us to do in this project considering the number of different features that will be available to us.

Tavakol (2016) explored this idea even further: again using a Linear Model, he used historical player data as well as historical results between the two teams going head to head in order to generate predictions. Due to the large number of features available, he used feature extraction and aggregation techniques to reduce the number of features to an acceptable level to train a Linear Model.

There are multiple ways to reduce the number of features to train a Machine Learning model: for instance, Kampakis used a hierarchical feature design to predict the outcome of cricket matches. On the other hand, in 2015, Tax et al. combined dimensionality reduction techniques with Machine Learning algorithms to predict a Dutch football competition. They came to the conclusion that they obtained the best results for the PCA dimensionality reduction algorithm, coupled with a Naive Bayes or Multilayer Perceptron classifier. It will be interesting for us to try different dimensionality reduction techniques with our Machine Learning algorithms if we have a large number of features we choose to use. This also shows us that a large amount of data might not be required to build a Neural Network model and achieve interesting results.

Bayesian networks have been tested in multiple different recent research papers for predicting football results. In 2006, Joseph built a Bayesian Network built on expert judgement and compared it with other objective algorithms, namely Decision Tree, Naive Bayesian Network, Statistics-based Bayesian Network and K-nearest neighbours. He found that he obtained a better model performance for the Network built on expert judgement, however expert knowledge is needed and the model quickly becomes out of date.

Another type of Machine Learning technique that has been used for a little longer is an Artificial Neural Network (ANN). One of the first studies on ANN was made by Purucker in 1996 to predict NFL games, who used backpropagation to train the network. One of the limitations of this study was the small amount of features used to train the network. In 2003, Kahn extended the work of Purucker by adding more features to train the network and achieved much better results, confirming the theory that Artificial Neural Networks could be a good choice of technique to build sports predictive models.

Hucaljuk et al. (2011) tested different Machine Learning techniques from multiple algorithm families to predict football scores: Naive Bayes (probabilistic classification), Bayesian Networks (probabilistic graphical model), LogitBoost (boosting algorithm), K-nearest-neighbours (lazy classification), Random Forest (decision tree), Artificial Neural Networks

They observed that they obtained the best results when using Artificial Neural Networks. This experiment is espacially interesting to us as we will want to test different algorithms in the same manner, to obtain the one that works the best for our data and features.

CONCLUSION

This study has presented a characterization and trends of development f machine learning techniques applied to prediction of soccer matches. Pitfalls to avoid while building predictive models have been highlighted and discussed.

REFERENCES

1. *H. Rue, O. Salvesen 2000. Prediction and retrospective analysis of soccer matches in a league. Statistician,*
2. *Crowder M., Dixon M., Ledford A., Robinson M., 2002. Dynamic modelling and prediction of English Football League matches for betting.*
3. *Forrest D., Simmons R., 2000. Forecasting sport: The behaviour and performance of football tipsters. International Journal of Forecasting.*
4. *Kuypers T., 2000. Information and efficiency: An empirical study of a fixed odds betting market. Applied Economics.*
5. *Goddard J., 2005. Regression models for forecasting goals and match results in association football. International Journal of Forecasting.*
6. *Hamadani B., 2006. Predicting the outcome of NFL games using machine learning. Stanford University.*
7. *Adam A., 2016. Generalised linear model for football matches prediction. KULeuven.*
8. *Tavakol M., Zafartavanaelmi H., and Brefeld U,. Feature Extraction and Aggregation for*

SJIF Impact Factor: 7.001| ISI I.F.Value:1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 5 | Issue: 10 | October 2020 - Peer Reviewed Journal

Predicting the Euro 2016. Leuphana University of Luneburg.

9. *Kampakis S., Thomas W., 2016. Using Machine Learning to Predict the Outcome of English County twenty over Cricket Matches. University College London.*
10. *Tax N., Joustra Y., 2015. Predicting The Dutch Football Competition Using Public Data: A Machine Learning Approach. Transactions on Knowledge and Data Engineering.*
11. *Joseph A., Fenton N.E., Neil M., 2006. Predicting football results using Bayesian nets and other machine learning techniques. Knowledge-Based Systems.*
12. *Purucker M.C., 1996. Neural network quarterbacking. IEEE Potentials.*