NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

PESS22016 - Creating an Expected Goal Model for Singapore football

Presented by Pathak Siddhant

Supervised by Assistant Prof Komar John

Introduction

Expected Goals (xG) is a widely used statistical tool in professional football for assessing team and individual player performance[1]. However, the creation of an accurate xG model is dependent on the quality and quantity of data used to train the models[2].

The aim of this study is to create an Expected Goal (xG) model for Singapore football, which predicts the probability of a shot taken resulting in a goal or not. We will treat this as a classic probability prediction problem.

Methodology

Our dataset contains temporal and spatial data on two seasons of English Premier League football matches, used to train models and compare their scores. To account for differences in data availability, we have split the training into **5 levels of information**. These levels are indicative of the amount of data provided to train the model.

Level 1	Level 2	Level 3	Level 4	Level 5
• Goals scored • Time • (x,y) coordinates • Home Team • Phase Type	• Goals scored • Time • (x,y) coordinates • Home Team	• Goals scored • Time • (x,y) coordinates	• Goals scored • (x,y) coordinates	• (x,y) coordinates

Chart 1: Levels of Information

The probabilistic model was trained on each level using various techniques, including Logistic Regression, K-Nearest Neighbours, Linear and No-Support Vector Classifiers, Decision trees, and a customised Deep Neural Network. The ensemble of models with **> 90% accuracy on the test set** was proposed after fine-tuning.

MODEL	LEVEL 1		LEVEL 3		LEVEL 5	
	Train Acc.	*Test Acc.*	*Train Acc.*	*Test Acc.*	*Train Acc.*	*Test Acc.*
Logistic Regression	0.716	**0.723**	0.714	**0.718**	0.584	**0.608**
Decision Tree	0.995	**0.992**	0.997	**0.994**	0.995	**0.992**
K-Nearest Neighbors	0.990	**0.985**	0.991	**0.986**	0.995	**0.993**
Linear SVC	0.722	**0.723**	0.714	**0.721**	0.603	**0.628**
No-Support SVC	0.906	**0.912**	0.907	**0.893**	0.908	**0.897**
Deep Neural Network	0.961	**0.9753**	0.945	**0.947**	0.988	**0.992**

Table 1: Evaluation Metrics for models across different levels (only 1,3 and 5 shown here)

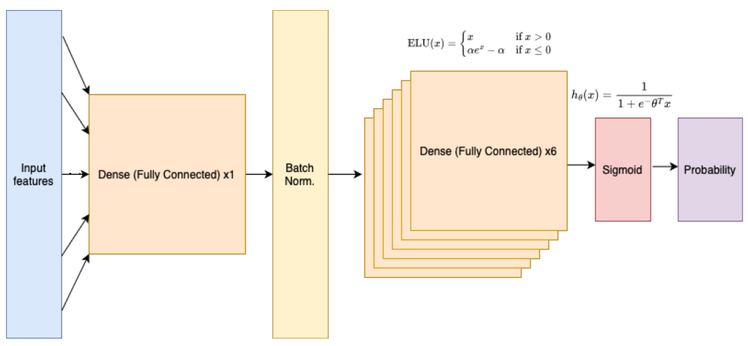

Fig 2: Deep Neural Network Architecture

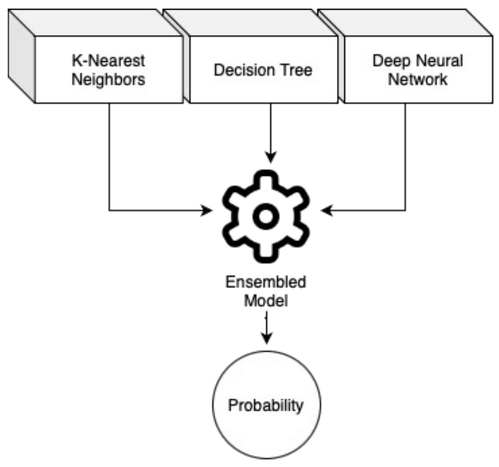

Fig 3: Proposed Ensemble Model

Exploratory Data Analysis

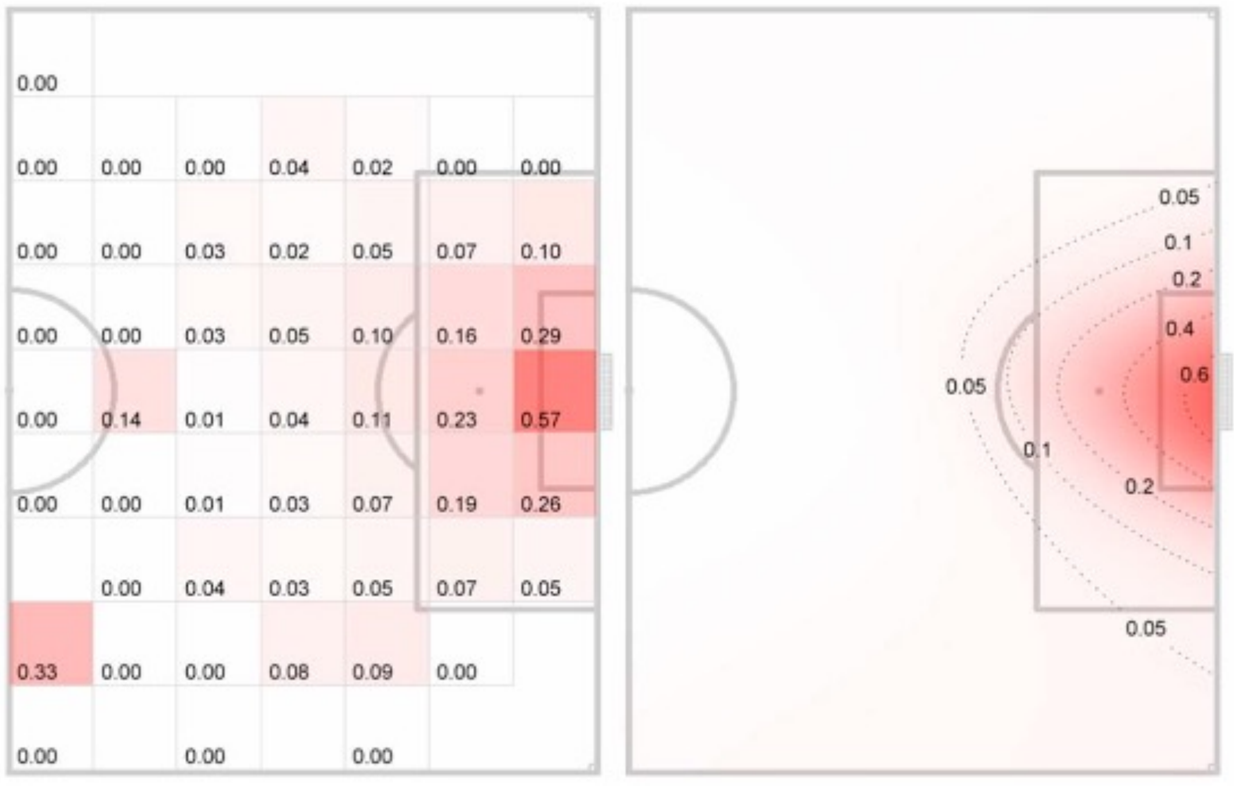

Fig 1: Observed (left) and estimated (right) shot conversion rates for open-play footed shots [3]

As shown in the above shot conversion rates, it is very important to note that the probability tends to zero as we go farther from the field. However, this statistical result does not consider other factors that have a physical and mental impact on the psyche of the player – namely, **time left, goals scored by both the teams, number of red cards/fouls, type of play (open play/set-piece)**. Regardless of the position, these dynamic natural features tend to influence the conversion rate as well.

Conclusion

Our research identifies accurate xG prediction models for Singapore football and similar contexts with limited data. Insights from various models contribute to the literature on xG modelling and machine learning in football, helping develop cost-effective methods to enhance performance analysis and decision-making in football. Our study enhances the literature on machine learning in football, with implications for future research in the field.

References

[1] Rathke, Alex. (2017). *An examination of expected goals and shot efficiency in soccer. Journal of Human Sport and Exercise.* 12. 10.14198/jhse.2017.12.Proc2.05.

[2] Scarf, Phil. (2006). *Modelling the outcomes of association football matches*. 48th Annual Conference of the Operational Research Society 2006, OR48. 59-72.

[3] Ruiz, H., Lisboa, P.J., Neilson, P., & Gregson, W. (2015). *Measuring scoring efficiency through goal expectancy estimation*. The European Symposium on Artificial Neural Networks.