Literature Review: European Soccer

Justin O'Donnell, Brian Pascente, Carson Holscher, Jacob Richardson

Shared to: henthorn@gmail

Teammate who read and summarized: Justin O'Donnell

Paper: Islam, K. T., & Nahid, M. H. (2024). Applications and implications of data-driven analytics in the football player valuation. *SHS Web of Conferences, 204*, 04006. https://doi.org/10.1051/shsconf/202420404006

Summary: In this paper, Islam, K. T., and Nahid H. analyze the applications and implementations of data-driven analytics in specific player evaluation. They do not propose a new model but instead use existing machine learning, artificial intelligence, and big data practices. The main machine learning models used in this paper are Random Forests, Gradient Boosting Machines, and Neural Networks. It talks about how these models can be used to become less subjective in decision making. The main aspects analyzed were the biometric statistics, market trends, and performance data. Additionally, it mentions how these models can be used in football management, including revenue optimization, fan behavior analysis, and tactical decision-making. It reviews how clubs use KPIs and performance analytics to identify undervalued talent and negotiate smarter transfer deals. Finally, the paper touches on model selection theory, optimization techniques, and the need for balanced data to improve valuation accuracy.

Applicability: In this paper, they focus on player valuation which is extremely applicable as it is currently being used by many teams in Europe. It also discusses the commonly used datasets and KPI's which are applied in the real world. These models can be used differently by each club depending on their strategies, but all can be used to take value from the data.

Issues: The main issues are soccer's naturally non-formulaic gameplay. This is why data use is more important in baseball than soccer. But it does not mention cleaning data, which will be a big part of our process. The paper says more "why" than "how" the analysis is conducted.

Teammate who read and summarized: Brian Pascente

Paper: Chandra B, Jennet Shinny D, Keshav Adhitya M (2024). Prediction of Football Player Performance Using Machine Learning Algorithm *Research Square* (https://doi.org/10.21203/rs.3.rs-3995768/v1)

Summary: This paper proposes a method utilizing machine learning algorithms to predict football player performance by analyzing historical data, considering key factors such as player statistics, team dynamics, and match conditions. The study uses data mining techniques to forecast match results by analyzing historical match data and identifying key features. Several classifiers including logistic regression, SVM, and Bayesian networks are used to test and refine the models. The system aims to predict player performance metrics like goals, assists, and defensive contributions, as well as predict injury probability and support player ranking and scouting. The project emphasizes meticulous feature selection and algorithm optimization to enhance accuracy, with the ultimate goal of providing actionable insights to coaches and team managers to optimize team performance and decision-making.

Applicability: This paper is applicable since it covers the same topic as our project.It outlines key stages such as data collection, preprocessing, and feature selection in order to get the best result, which are ideas we can use for our topics

Issues: This paper didn’t go too much into detail on justifying the specific algorithms for the project. Between logistic regression, SVM, and Bayesian networks, the reasoning behind the superior performance of certain classifiers (if any were found to be significantly better) is not thoroughly discussed.

Teammate who read and summarized: Carson Holscher

Paper: Predicting the value of football players: machine learning techniques and sensitivity analysis based on FIFA and real-world statistical datasets (link)

Summary: This paper discusses its method for predicting soccer player trade value using machine learning & details the models it chose & why. (The author combined svr, rfr, xgb, & dtr models using Dempster-Shafer theory, & optimized with sand cat swarm). The system predicts the player's value based on factors like age, performance metrics, reputation, & attributes like height, weight, & dominant foot. The paper spends a great deal of time discussing the mathematical background of each of the models it looked into, along with detailing the algorithms for sand cat swarm optimization & hunger games search optimization in order to give the reader an understanding of why the authors chose to do what they did.

Applicability: This paper is applicable since it is about the same thing as our project. It mainly discusses what models it used to make its predictions & the reasoning behind those choices. The paper also makes mention of a fifa database & where to get ahold of it.

Issues: The paper doesn't talk very much about data collection or preparation. It just assumes that you already have the information & want to operate on it.

Teammate who read and summarized: Jacob Richardson

Paper: Predictive Analysis and Modelling Football Results Using Machine Learning Approach for English Premier League (link)

Summary: Predicting the outcome of sports is a highly sought-after capability and is nearly impossible with machine learning and data mining techniques due to the unpredictable and versatile nature of human capability. Approaches to this issue have been primarily in two categories: result-based and goal-based studies. Result-based studies (the one used in the write-up) aim to predict result class (away win, home win, or draw), while goal-based studies predict the number of goals for each team in a match. Some predictors use datasets such as player stats, while others use weather, injuries, or even psychological states.

Applicability: The paper explores many different techniques for predictive models, the advantages of certain datasets, and overall effectiveness of the predictions. We can learn from many others who have had the same objective as we do for this project.

Issues: Slight lack of dataset insight and limited code/feature engineering detail.