Predicting the impact of soccer transfers on team performance using neural networks

Matthew Aquilina
Mentored by Kieran O'Connor

Science and Mathematics Academy

Booz | Allen | Hamilton®

Introduction

Each year, soccer clubs invest millions of dollars in player transfers, yet many signings fail to improve team performance. In response, the Union of European Football Associations (UEFA) introduced financial fair play regulations to limit financial losses (Barron et al., 2018). These regulations forced clubs to be more selective when recruiting players. As a result, scouting departments increasingly rely on key performance indicators (KPIs) to evaluate players. For example, the pressing index measures the average speed after losing possession. Data for these KPIs are widely available through open-source providers such as Opta Sports.

Beyond evaluating team success, previous studies have applied machine learning algorithms to predict the outcome of soccer transfers. One study used a classification approach, defining a successful transfer as a player achieving an assessment score above a predetermined threshold (Ćwiklinski et al., 2021). This score is calculated using a weighted average of psychological, physical, and technical features.

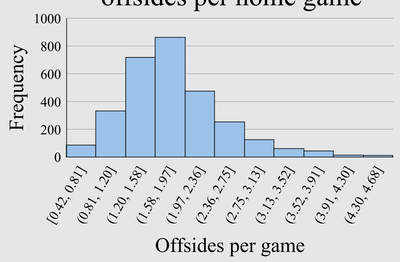
The goal of this project is to develop an algorithm that predicts the impact of a soccer transfer on team performance. The analysis focuses on data from five major leagues, consistent with prior research such as Pappalardo et al. (2019). The English, German, Spanish, Italian, and French top divisions were selected due to their comparable skill levels and the availability of extensive public data.

Methods and Materials

A player's contribution was defined as the change in a team's goal difference (GD) between when the player was on and off the field.

This project was developed in Visual Studio Code using the Python programming language. Open-source data were collected from the website Kaggle (Cariboo, 2025). A combination of player and team statistics from 2015 to 2022 was used to train the neural network. Missing values were imputed using the k-nearest neighbor algorithm.

Distribution of opponent offsides per home game



Additional features, such as shot accuracy, were derived from the original variables. Histograms and regression analysis were used to determine which features to include in the neural network (Graph 1).

Graph 1 (left): The distribution of average offside offenses an opponent commits while at the selected team's home stadium is shown. There is a relatively small range, so this variable is not useful in model training.

Methods and Materials (continued)

Ultimately, nine player variables, nine team variables, and nine positions were selected to train the neural network. The network was composed of two linear layers and trained using a learning rate of 0.003, which defines how fast the network's weights converge. The parametric rectified linear unit activation function was used to introduce non-linearity to the network, enabling the identification of complex patterns. A graphical user interface was also developed (Figure 1), which allowed users to input the selected features across three different input tabs. The predicted player contribution is then displayed on the final tab after clicking a button (Figure 2).

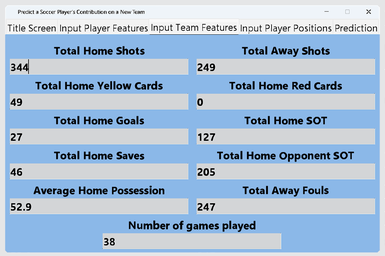


Figure 1 (left): The user entered all required team features into the eleven input boxes.

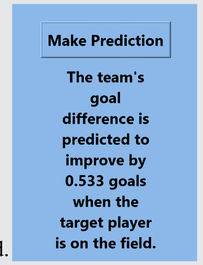
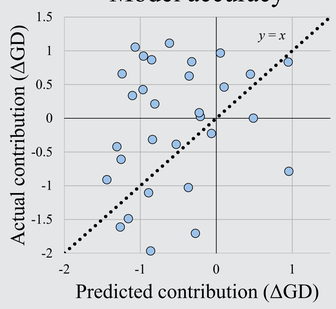


Figure 2 (right): After pressing the button, the player's predicted contribution is displayed.

Results

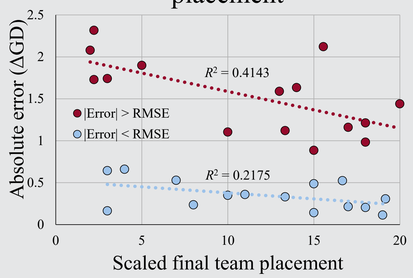
A paired *t*-test revealed that players ($n = 30$) were predicted to have a larger negative contribution (ΔGD) to their team ($M = -0.57$, $SD = 0.69$) than they actually contributed ($M = -0.083$, $SD = 0.91$), $t(29) = -2.5$, $p = .018$; 95% CI [−0.89, −0.089] (Graph 2). The neural network resulted in a root mean squared error (RMSE) of 0.67, with the greatest error occurring when the player joined an already successful team (Graph 3).

Model accuracy



Graph 2 (above): Players with their contribution predicted perfectly would fall on the $y = x$ line. Points at the top left were underestimates.

Error by team's final placement

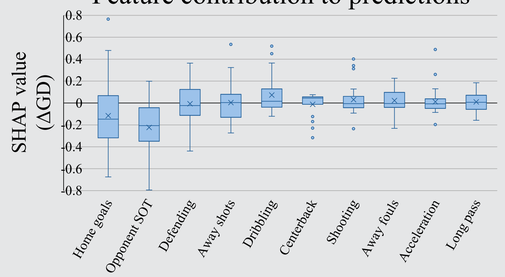


Graph 3 (above): Team placement was scaled out of 20, with error grouped relative to the RMSE. Absolute error is greatest for teams who finished higher in the standings.

Results (continued)

The Shapley Additive exPlanations (SHAP) values were calculated for each feature in the neural network. SHAP values indicate how much each feature contributed the overall prediction (Graph 4).

Feature contribution to predictions



Graph 4 (left): The distribution of SHAP values for the ten most influential features in the model (left to right) are shown. Wider distributions indicate greater importance. The number of home goals had the greatest contribution. Of those displayed, long pass accuracy had the least contribution to the prediction.

Conclusion

The neural network produced accurate predictions on transfers from the same period as the training data but performed worse on more recent transfers, indicating limited generalizability over time. Further analysis revealed that prediction error was highest for players joining an already successful team.

The number of home goals scored by the target team was the most influential variable, with higher values reducing predicted player contribution. This is likely because players had less opportunities to improve stronger teams.

Future studies can investigate if the addition of more leagues would alter the most influential variables in the network, as well as determining if a similar framework can be applied when selecting players to represent a national team.

References

Barron, D., Ball, G., Robins, M., & Sunderland, C. (2018). Artificial neural networks and player recruitment in professional soccer. *PLoS ONE*, *13*(10), e0205818. https://doi.org/10.1371/journal.pone.0205818

Cariboo, D. (2025). *Football data from Transfermarket* [Data set]. Kaggle. https://www.kaggle.com/datasets/davidcariboo/player-scores

Ćwiklinski, B., Giełczyk, A., & Choraś, M. (2021). Who will score? A machine learning approach to supporting football team building and transfers. *Entropy*, *23*(1), 90. https://doi.org/10.3390/e23010090

Pappalardo, L., Cintia, P., Ferragina, P., Massucco, E., Pedreschi, D., & Giannotti, F. (2019). PlayeRank: Data-driven performance evaluation and player ranking in soccer via a machine learning approach. *ACM Transactions on Intelligent Systems and Technology*, *10*(5), 1–27. https://doi.org/10.1145/3343172