Data-Driven Valuation of Football Players Using Machine Learning Algorithms

Degree programme: BSc in Industrial Engineering and Management Science
Thesis advisor: Prof. Dr. Jörg Grimm
Expert: Tim Luginbühl
Industrial partner: Berner Fachhochschule BFH, Biel/Bienne

The competitive landscape of the football industry often leads clubs to prioritize sporting success over economic viability, resulting in financial strain. On-field shortcomings are often quickly addressed with new transfers, leading to potential losses and inability to invest enough for the next transfer period. Using machine learning models to determine the optimal timing for player transfer will be crucial, as this timing significantly impacts the club's financial health.

Introduction and Objectives

Football clubs must integrate economic concepts into player transfer decisions to maintain finances in a healthy, sustainable manner. The TCO Model (Total Cost of Ownership, Fig. 2) is an ideal approach, treating players as club assets impacted by lifecycle factors. However, a key limitation in football is forecasting players' remaining market value to operationalize TCO. This study identifies factors influencing player market estimation and integrates them into machine learning models for TCO framework implementation.

Research Design

To forecast the market value of football players, data to train and test the model was collected through web crawling, following theoretical background research to identify influential variables such as performance metrics (from match reports on FBREF.com), player age, position, and injury susceptibility (from Transfermarkt). A Random Forest Regressor was used to forecast the market value of football players within different groups like keeper, defender, midfielder, and forward, which were additionally divided according to their respective ages. Top features for each cluster were identified to gain insights into which metrics to look out for to reliably predict the market value of a given player.

Results

The performance of the random forest regressor (Fig. 1), measured by RMSE (Root Mean Square Error) and MAPE (Mean Absolute Percentage Error), indicates that it generally performs well, with lower RMSE values suggesting better accuracy in predicting market values for specific groups like aging defenders (RMSE = 0.98) and midfielders (RMSE = 1.68). However, the model struggles with younger and prime forwards, as evidenced by higher RMSE values (14.4 and 6.04, respectively) and corresponding high MAPE values. For instance, predictions for young forwards must be considered within an approximate range of ±14.4 million euros around the predicted value or 80% deviation from the actual values. Age, minutes played, and different passing metrics have emerged as top features.

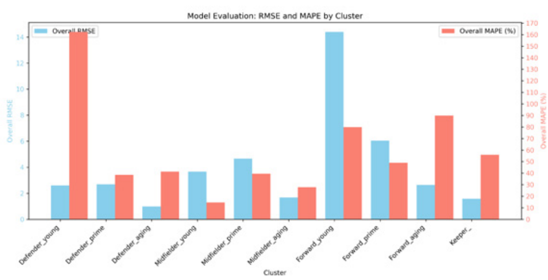

Figure 1: Forecast accuracy regarding the overal RMSE and MAPE of each cluster

Implications and Recommendations

Based on the model results, the model can predict the market value of the football players by working with different performance metrics, injury susceptibility and player age itself. But to reduce the error margins in the market value forecast especially for the younger players, it is recommended to implement the status- and popularity variables of the respective players for the training phase of the model. Since the model tries to approximate the market value prediction with the values from Transfermarkt, it is necessary to cover those two groups of features as well to increase the forecast accuracy.

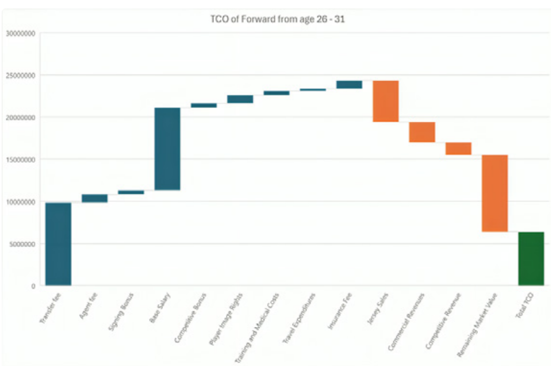

Figure 2: Use Case of the implementation of model forecasts in the TCO concept

Vakeesan Thayaladevan
Business Engineering