TU WIEN Informatics

Recommending Similar Players using Technical Performance Indicators

Alexander Lorenz
Master Data Science

WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS

Wirtschaftsuniversität Wien
Institute of International Business
Department of Global Business and Trade
Supervisor: Assoz. Prof. PD Dr. Jakob Müllner
Contact: jakob.muellner@wu.ac.at

TU WIEN TECHNISCHE UNIVERSITÄT WIEN

TU Wien Informatics
Institute of Information Systems Engineering
Data Science Group
Co-Supervisor: Univ.Prof. Dr. Allan Hanbury
Contact: allan.hanbury@tuwien.ac.at

1. Motivation

- Modern football has become increasingly **data-driven**, particularly in **game analysis**, **physica health monitoring**, and **player scouting**
- Yet, clubs still face the **sudden loss of key players** due to transfers or long-term injuries, which weakens team performance and **demands replacements** who match not only the position but also the playing style of the missing player
- However, **finding ad-hoc replacements** among thousands of candidates remains **highly challenging**, as **scouts** rely on statistics, visualizations, and video clips but are **constrained by human capacity** to a limited pool of players
- Meanwhile, **transfer fees** and player salaries represent some of the **largest cost factors** for professional clubs, and **misjudgments** in recruitment can **lead to significant financial losses**

2. Objective

The goal is to develop a data-driven recommendation method that instantly identifies top-k player replacements based on advanced statistics that captures player characteristics.

3. Data

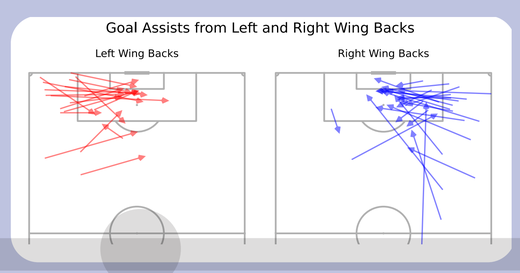
- The StatsBomb Python API serves as the data source source
- It provides **detailed game-event data** for the 2015/16 season across the Premier League, Bundesliga, La Liga, Serie A, Ligue 1, and the UEFA Champions League
- The raw dataset contains of **6,391,338 rows and 122 columns** where each row records an event with its type, outcome, and location

4. Feature Engineering

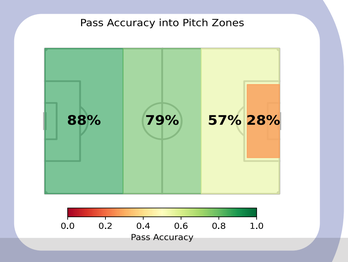
Aggregated Feature Vector
- Raw event-level data transformed into aggregated player vectors to derive general statistics
- Contextual variables added that refine generic statisitcs by accounting for pitch location, under pressure situation, and its outcome
- 1000+ features generated, grouped into dimension goalkeeping, defending, possession, passing, and shooting

Heatmap
- evaluated for each feature dimension
- Reduced to a fixed set of principal components

5. Exploratory Data Analysis



- Extracted features set compromises **3,069 players and 1,032 features**
- **Distinct movement patterns** across positions show varying involvement during matches
- **Role-based** action **patterns** emerge
- Performance **metrics**, such as pass accuracy, **vary with pitch location**

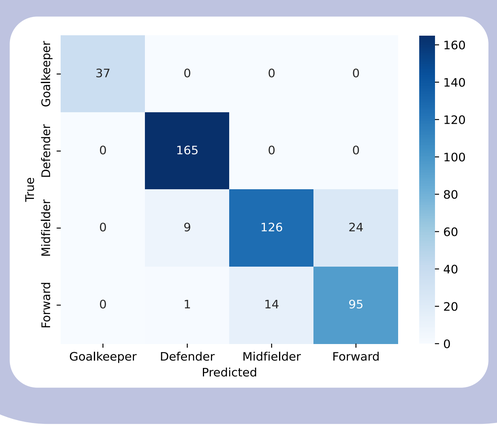


6. Feature Selection

- Reducing features is necessary to **address multicollinearity** and prevent **overfitting**
- Two feature sets were created and evaluated
- **Feature Set 1**: Logistic Regression with L1 regularization
- **Feature Set 2**: Prefiltering correlated features, followed by manual selection using domain knowledge

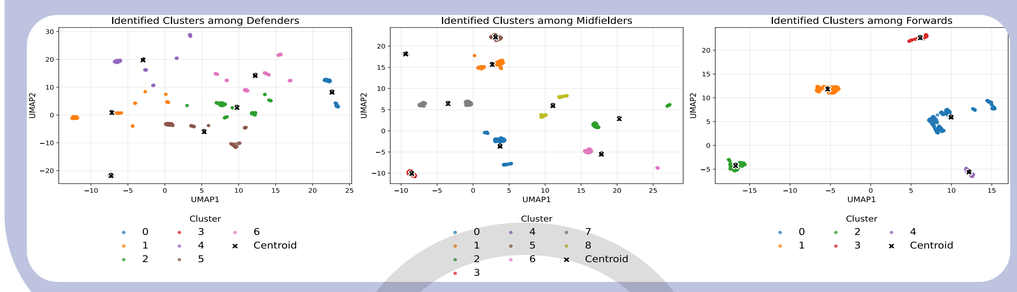
7. Modeling

- **Set of experiments** configurations with selected models evaluated predictive power using 5-fold cross-validation to **predict player positions**
- **Best model** chosen based on **stability** (lowest accuracy std) and **performance** (highest mean accuracy)
- LightGBM **achieved 90% accuracy**, with **perfect** prediction **for goalkeepers** and **defenders**



8. Clustering

- Unsupervised **K-Means applied** on SHAP-transformed features, reprojected **to 2D** with UMAP
- **At positions level**, clusters achieved **perfect homogeneity** and **completeness score** (goalkeeper, defender, midfielder, and forward)
- **Within positions**, clusters were more **heterogeneous**, grouping diverse role labels



9. Recommendation

- Split feature set into **test collection** and **player database**, stratified by position
- **Generated recommendation lists** (k=10, k=30) using cosine similarity; relevance defined by matching position labels
- Results show **accurate recommendations**, with **relevant players ranking high**

	AP@10	MAP@10	MRR@10	AP@30	MAP@30	MRR@30
$Goalkeeper$	1	1	1	1	1	1
$Defender$	0.7	0.93	0.95	0.66	0.93	0.95
$Midfielder$	0.9	0.80	0.85	0.93	0.80	0.85
$Forward$	0.8	0.84	0.88	0.6	0.84	0.88

10. Conclusion

- **Contextual** and **spatially-aware** aggregated **player vectors** provide a **strong representation of** positions by encoding **on-field behavior**
- **Player** vectors **reveal heterogeneous clusters** within positions, **indicating** that **players with different** official **role labels** may **share overlapping on-field** performance **patterns**
- **Engineered features** combined with cosine similarity **enable fast, good-quality scouting shortlists**