CAPSTONE PROJECT
PREDICTING TRANSFER SUCCESS

Rishvanyas Premanand, Balint Farkas, Laurenz Flender, Kevin Drost

COMPANY

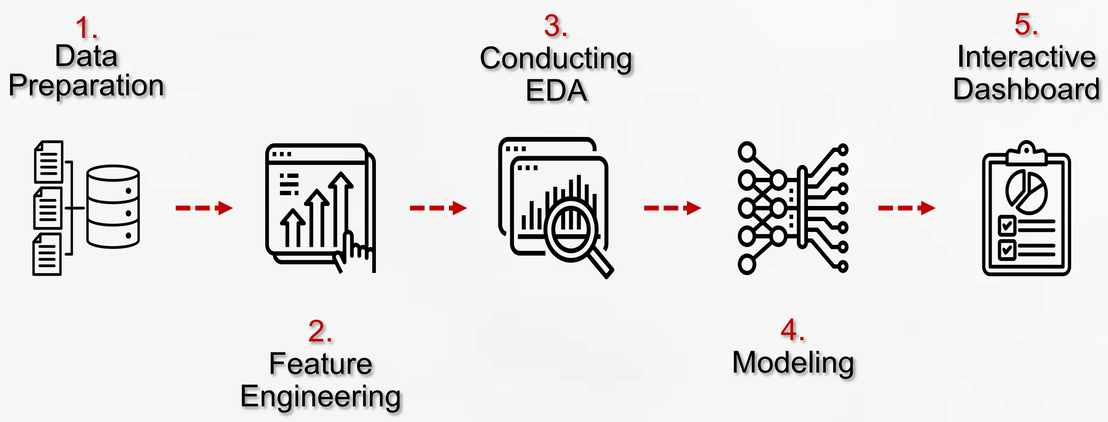
1. FC Köln is a historic football club founded in 1948. As of the 2023/24 season, the club boasts around 140,000 members, €159 million in revenue, and a profit of €11.8 million. The club faced a transfer ban in 2024, which limited its ability to sign new players. The club's transfer strategy plays a key role in ensuring both sporting success and financial stability.

With a new focus on data-driven decision-making, the scouting department seeks to modernize its approach to identifying and evaluating potential signings. This project aims to support the scouting team by providing a predictive model to enhance recruitment decisions and increase the likelihood of successful transfers.

CASE

The club faces the challenge of signing new players who can contribute both on the pitch and financially. Our goal of this project is to predict whether a potential new signing will be successful in their first season after transfer. A successful signing is defined as a player who plays at least 50% of the possible minutes in a season.

METHODOLOGY



1. Merging transfer-related datasets into a unified dataset, handling missing values, duplicates, and standardizing formats.
2. Transforming of raw data into meaningful features, such as performance metrics, market value indicators, and club-related variables.
3. Identifying key patterns and relationships between player characteristics and transfer success.
4. XGBoost, Logistic Regression, and more models to predict transfer success, using post-transfer playing time as the target variable.
5. The best-performing model is integrated into an interactive dashboard to support decision-making on transfer success.

DATA

To build the model, we integrated a variety of data sources:

- Transfers: Transfers from 2019 to 2024 including transfer fees
- Market Value: Transformed market values from 2019 to 2024
- Player Information: Position, height, preferred foot, and more
- Performance Statistics: Detailed lineups from all competitions (2019–2024) to calculate playing percentage, goals and assists per minute and clean sheets

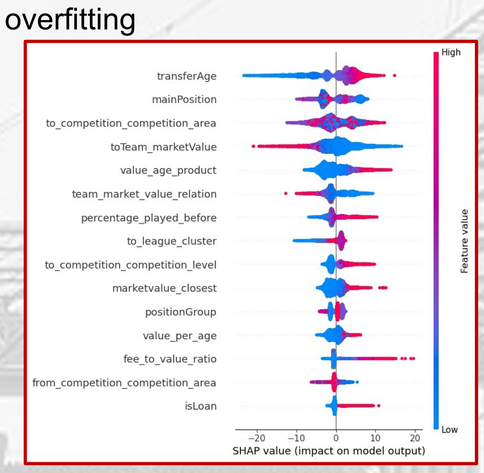
The final merged dataset contained a wide range of features, resulting in a high-dimensional structure. Additionally, extensive missing values, particularly for players in lower-tier leagues with limited data coverage, posed a significant challenge. This highlights our need for a model that can effectively handle incomplete and complex data, ensuring adaptability across diverse leagues and suitability for robust transfer success prediction.

MODELS

Three models were tested to tackle the data challenges and predict transfer success:

- Logistic Regression as a baseline linear classifier
- Histogram Gradient Boosting captures non-linear patterns efficiently
- XGBoost offers strong regularization to reduce overfitting

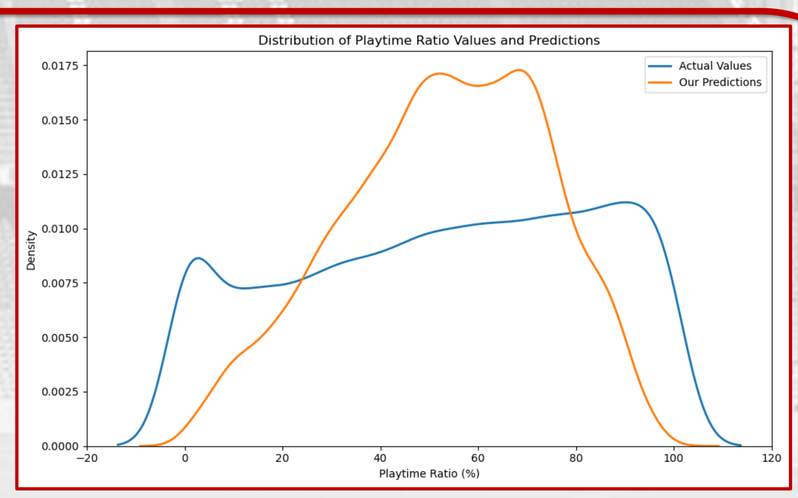
For XGBoost, prior season playing time is important but not the top factor. Features such as transfer age, main position, and destination league or team characteristics have greater influence, indicating that the model accounts for league or team strength and player context rather than relying solely on past playing time.



RESULTS

Model Comparison:
Logistic Regression showed better recall for successes, meaning it identified a higher proportion of successful transfers. Histogram Gradient Boosting achieved higher precision for failures, making it more reliable at correctly identifying unsuccessful transfers. XGBoost balanced both precision and recall, providing more robust predictions across classes.

Key Takeaway:
XGBoost was chosen as the final model for its strong performance, ensuring the most accurate predictions for scouting decisions among our models. The plot on the right shows the distribution of predicted versus actual playing percentages. The model predicts fewer extreme values due to aggregation bias, where predictions cluster towards the mean. As a result, very low or high playing percentages are less frequent in the predicted distribution. Even with added features and regularization, fully matching the true distribution remains challenging with the available data.



The interactive dashboard integrates the trained model, allowing users to input player stats, performance data, and attributes to simulate any real or conceptual player. It then predicts the expected playing percentage for their first season post-transfer, supporting data-driven scouting decisions. Scan the QR code above to try out the dashboard and test predictions directly.

RECOMMENDATION

To further improve model performance, integrating more granular, position-specific metrics such as passing accuracy, tackles, interceptions, and aerial duels is recommended. Collecting team-specific data would also allow training models tailored to each club's needs, enhancing practical relevance.
It is important to note that player archetypes and positions differ in expected playing time. For example, younger developing players are not meant to play as many minutes as starters. Thus, defining transfer success purely by playing percentage can be limiting, and incorporating role-adjusted success definitions would improve model validity.
Finally, we explored the data and used our model insights to develop practical guidelines for making better transfer decisions.

Our 4 Do's and Don'ts for Transfer Decisions:

- Do loan young players (18–22 years) to support development
- Do sign players who played a lot of minutes in previous seasons
- Don't loan out players over 22 years old, as it offers limited benefit
- Don't sign older, experienced players if you are a low market value team