Laurenz Flender, Kevin Drost, Rishvanyas Premanand, Balint Farkas

Capstone Project - 1. FC Köln

Internal Supervisor: Christopher Coors

July 17, 2025

Eidesstattliche Versicherung

Hiermit versichere ich an Eides statt, dass ich die vorliegende Arbeit selbstständig und ohne die Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe. Alle Stellen, die wörtlich oder sinngemäß aus veröffentlichten und nicht veröffentlichten Schriften entnommen wurden, sind als solche kenntlich gemacht. Die Arbeit ist in gleicher oder ähnlicher Form oder auszugsweise im Rahmen einer anderen Prüfung noch nicht vorgelegt worden. Ich versichere, dass die eingereichte elektronische Fassung der eingereichten Druckfassung vollständig entspricht.

Die Strafbarkeit einer falschen eidesstattlichen Versicherung ist mir bekannt, namentlich die Strafandrohung gemäß § 156 StGB bis zu drei Jahren Freiheitsstrafe oder Geldstrafe bei vorsätzlicher Begehung der Tat bzw. gemäß § 161 Abs. 1 StGB bis zu einem Jahr Freiheitsstrafe oder Geldstrafe bei fahrlässiger Begehung.

Laurenz Flender, Kevin Drost, Rishvanyas Premanand, Balint Farkas

Köln, den 18.07.2025

Contribution Statement

Team Members

1. Laurenz Flender
2. Kevin Drost
3. Rishvanyas Premanand
4. Balint Farkas

Contribution Statement

Team Member	**Contribution Description**
Laurenz Flender, 7430935	Equal Contribution
Kevin Drost, 7434810	Equal Contribution
Rishvanyas Premanand, 7434745	Equal Contribution
Balint Farkas, 7430884	Equal Contribution

Abstract

In the high-stakes environment of professional football, clubs face high pressure to make data-informed decisions that balance sporting success with financial sustainability. This project addresses the challenge of predicting transfer success for players entering domestic leagues. Focusing on 1. FC Köln and transfers within European football, the project analyzed over 70,000 transfer records between 2019 and 2024 to develop a machine learning pipeline to forecast post-transfer playing time. Exploratory Data Analysis (EDA) revealed key trends, such as the overperformance of players from Japan, Switzerland, and Portugal's Liga Portugal, and identified ideal loan destinations in Spain and Belgium. A refined XGBoost model demonstrated moderate predictive power ($R^2 = 0.28$; RMSE = 25.8), and was deployed via an interactive dashboard designed to support scouting and recruitment decisions. Results reveal that the player's age was the most impactful feature, where older players are predicted to play more in the first season after transfer, peaking at the age of 33. The tool enhances transparency, supports risk assessment, and integrates seamlessly with the strategic objectives of a modern football club.

Contents

List of Figures

List of Tables

1 Introduction

The professional football industry represents a high-stakes domain where both sporting success and financial sustainability are tightly interlinked. Clubs must make strategic decisions about player recruitment, match performance, and financial planning while operating in an increasingly competitive and regulated environment. The European football transfer market has become one of the most visible and financially impactful mechanisms in the sport, with the mobility of players widely accepted as a natural and integral part of the game (Szymanski, 2010).

Within this domain, the overall business problem is the uncertainty and risk associated with player transfers. Transfer fees are rising, player performance remains difficult to predict, and financial missteps can result in long-term setbacks or even sanctions. Inaccurate or overly speculative recruitment can strain finances and diminish on-pitch performance, especially for mid-sized clubs like 1. FC Köln (Barajas & Rodriguez, 2010).

This project is in partnership with 1. FC Köln, a traditional German football club with around 140,000 members (1. FC Köln, 2024a) and €159 million in revenue in the 2023/24 season (1. FC Köln, 2024b). The club faced a transfer ban in 2024, which significantly limited its ability to sign new players and emphasized the importance of effective, forward-looking transfer planning (Kicker, 2023). In response, the club's scouting department has adopted a data-driven recruitment strategy to reduce uncertainty in transfer decisions.

The concrete research question addressed by this project is: Can the success of a new player transfer within the European football market be predicted in their first domestic league season, based on historical and contextual data available at the time of transfer?

Success is defined as playing at least 50% of possible minutes in the first season post transfer, excluding injury spells. This serves as a tangible performance proxy that reflects tactical trust from the coaching staff, and imme.

To answer this question, a predictive machine learning model was developed and trained on historical player transfer data. The model incorporates player attributes, transfer details, and past performance indicators to estimate the likelihood of success. Its output supports scouting departments in identifying high-potential signings before financial commitments are made.

This approach contributes to the broader business problem by improving transparency in transfer-related decision-making, quantifying individual player risk and upside, and supporting compliance with financial constraints and long-term sustainability objectives.

2 Method

This chapter outlines the methodological framework applied to predict the success of football player transfers. As illustrated in Figure 2, the project progressed through several overlapping phases between April and July. Notably, a scope change in June extended the dataset and introduced new requirements for the final application, which led to the inclusion of additional modeling techniques and a stronger focus on interpretability. The process consists of five key steps (see Figure 1):

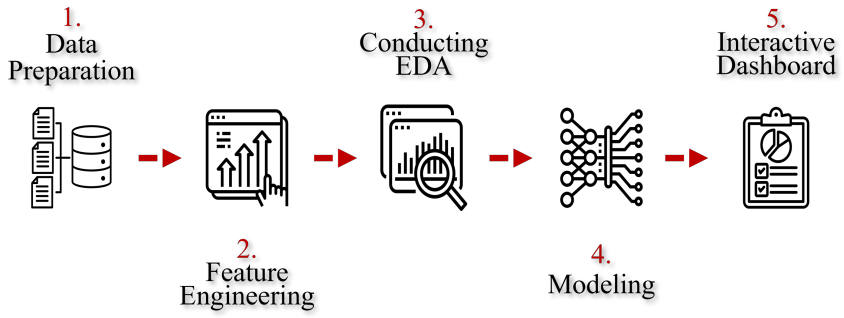

Figure 1: Methodology of the project

It begins with the integration of multiple datasets into a unified structure, addressing missing values, duplicates, and inconsistencies across sources. Subsequently, raw data is transformed into a comprehensive set of features, capturing performance metrics, market value indicators, and contextual club information. These enriched inputs enable the identification of relevant patterns and relationships between player attributes and transfer outcomes. In the modeling phase, several algorithms were applied, including Logistic Regression and Histogram-based models. XGBoost was ultimately used to optimize predictive performance regarding post-transfer playing time. Finally, the best-performing model is embedded into an interactive dashboard to support data-driven decision-making in player recruitment.

2.1 Data Compilation and Preprocessing

To build a predictive model for transfer success, a comprehensive dataset was compiled by merging various sources from *transfermarkt.de*. The site provides community-validated player information, including transfer details, match appearances, and market value estimates (Prockl & Frick, 2018). Due to this structure, the initial raw database was considerably larger than the final analytical dataset but required extensive cleaning and preprocessing. The final dataset integrates the following domains:

- **Transfer data:** origin and destination clubs and leagues (*fromTeam*, *toTeam*, *from_competition_level*, *to_competition_area*)

- **Market values:** player and team valuations from 2019–2024; used in raw form for modeling and log-transformed for visualization
- **Player attributes:** *height*, preferred *foot*, *mainPosition*, and *positionGroup*
- **Performance statistics:** full lineup data from all matches between 2019 and 2024

Before modeling, rows with missing target values, infinite values, or implausible inputs were removed. Numeric columns containing infinite values were clipped to the maximum or minimum finite value. Duplicate entries were dropped. Rare categorical values (e.g., unusual clubs or nationalities) were grouped under `other`.

After filtering and merging, the resulting data set contains around 70,000 transfer records from the period 2019-2024. It maintains structural diversity across leagues, competitions, and player profiles, offering a solid foundation for machine learning.

2.2 Feature Engineering

To improve model performance, a range of relevant features were developed to reflect player performance, situational context, and market valuation. An overview of all variables is provided in Table 1, while Figure 3 visualizes the distributions of selected features.

A central indicator in the dataset is *percentage_played*, which measures the share of minutes a player spent on the pitch relative to the total possible minutes, calculated as matchday squad appearances multiplied by 90. This metric was computed for both the season before and after the transfer, based on complete lineup and match records. Building on this measure, the binary outcome variable *success* was defined: a value of 1 indicates that a player featured in at least 50% of the possible minutes in the season following the transfer. The variable *success_before* applies the same logic to the preceding season and serves as an indicator of pre-transfer reliability. Further variables were derived from the data as follows:

- **Performance-based features:** *scorer_before_grouped_category* (binned goals and assists for midfielders and attackers) and *clean_sheets_before_grouped* (binned clean sheets) were computed for the pre-transfer period. These features were derived through extensive aggregation of individual match and lineup data across all games. A clean sheet was defined as a match in which the player was on the pitch for at least five minutes and their team did not concede any goals.
- **Market valuation indicators:** *value_per_age* and *value_age_product* (interactions between player age and value), and *team_market_value_relation* (ratio of destination to origin club value to model transfer direction).
- **Contextual transfer variable:** *foreign_transfer* indicates whether the player moved to a club in another country. This feature captures cultural and tactical adaptation challenges, which are empirically shown to affect transfer performance (Jarjabka, Fűrész, & Havran, 2024).

All categorical features (e.g., *foot*, *mainPosition*, *positionGroup*) were encoded using `pandas.Categorical` to support native handling in tree-based models.

2.3 Models

To predict transfer success, several machine learning models with varying levels of complexity were implemented.

2.3.1 Logistic Regression

As a baseline, a logistic regression model was employed to predict the binary outcome variable *success*. Due to its interpretability and suitability for binary classification tasks, logistic regression was selected as an initial approach. The model produced a recall of 0.83 and a precision of 0.70 for successful transfers, resulting in an overall accuracy of 69% (see Figure 2). For non-successful transfers, the recall was 0.47.

While appropriate for initial classification, logistic regression treats transfer performance as a binary outcome. This may not fully reflect the continuous nature of playing time. Although success was initially defined as playing at least 50% of possible minutes (see Section 1), the binary threshold oversimplifies the underlying variation in playing time, which ultimately led to a refinement of the modeling scope towards a continuous target variable. In addition, logistic regression does not inherently capture complex interactions between variables such as age, position, and club strength. To account for these aspects, the following section introduces boosted tree models, which allow for more flexible modeling of transfer outcomes.

2.3.2 Boosted Trees

The following algorithms builds trees on binned versions of continuous features, grouping values into discrete intervals before training. Binning not only reduces memory usage and speeds up computation, but also helps stabilize model behavior when variables have skewed or noisy distributions. Given the presence of variables such as age, market value, and prior playing time, histogram-based gradient boosting offered an efficient and interpretable way to model non-linear relationships in the data(Scikit-learn, 2024).

While this method provided a strong baseline, it lacked the fine-grained regularization and control over tree construction necessary to manage the complex interactions and heterogeneity present in the dataset. The HistGradientBoosting model achieved a reasonable R^2 of 0.226 and an accuracy of 0.68, as shown in Table 4, indicating moderate predictive power in a behavior-driven domain. As a result, the modeling approach was extended to XGBoost, which offered greater flexibility and performance through advanced boosting techniques(Chen & Guestrin, 2016).

To achieve higher precision, the XGBoost library (Extreme Gradient Boosting) was especially advantageous for several reasons. It is built on gradient-boosted decision trees, and combines a series of weak learners into a strong predictor, enabling the model to capture non-linear relationships in the data (Chen & Guestrin, 2016; Chen & contributors, 2025). Additionally, it provides extensive options for regularization, which help mitigate overfitting, particularly important given the high-cardinality and noisy features present in the dataset (Chen & contributors, 2025).

Root Mean Squared Error (RMSE) was used as the loss function, as it is standard for continuous targets and the distribution of the target variable did not exhibit zero inflation or sufficient skewness to justify alternative loss functions (Chai & Draxler, 2014). To further improve generalization, the model was trained with 5-fold cross-validation.

One key challenge in working with XGBoost is tuning its hyperparameters for optimal performance. To address this, Bayesian optimization was employed using the scikit-optimize library (Snoek, Larochelle, & Adams, 2012; Head, MechCoder, et al., 2021). The process began with a broad search space and was later refined to narrower intervals based on the most promising results. Table 3 displays the final hyperparameter values used.

The resulting model used 1083 estimators with a learning rate of 0.0190. This configuration allows the model to learn gradually over many rounds, effectively capturing complex feature interactions without overfitting. A maximum tree depth of 6 was chosen to ensure the trees could model important hierarchical relationships, such as how market value depends on age and position, while remaining generalizable. Other hyperparameters were selected to minimize the gap between training and validation accuracy (see Chapter 3).

This configuration proved more robust than the alternative models tested. As a result, the final predictions were generated using the XGBoost model without any stacking or ensembling. However, for interpretability, a LightGBM (Light Gradient Boosting Machine) model with similar predictive performance was also trained. This model was preferred for feature attribution analysis, as it provided clearer insights into variable importance and model behavior(Lundberg & Lee, 2017).

3 Results

Next, the results are explored across three dimensions: the underlying patterns in the data, the performance of the predictive model, and the interactive dashboard designed to support scouting decisions. Together, these components offer insight into player characteristics, predictive model performance, and the practical application of predictions in a scouting context.

3.1 Descriptive Results

A key objective of this project, as prompted by 1. FC Köln, was to uncover patterns and recent trends related to successful player transfers into the Bundesliga and outbound loans from German football clubs. To support the scouting department of 1. FC Köln in making informed decisions, a focused Exploratory Data Analysis (EDA) was conducted. This analysis aimed to identify promising markets for recruitment as well as optimal destinations for player development through loans.

The EDA initially focused on the nationalities of incoming transfers to the Bundesliga, where success was measured by the percentage of minutes played post-transfer, serving as a proxy for immediate impact. Among all nationalities examined over the past five years, Japanese and Swiss players demonstrated particularly strong outcomes, averaging 67.6% and 61.3% of total minutes played, respectively (see Figure 4). These findings suggest that players from Japan and Switzerland tend to integrate well and contribute meaningfully upon arrival in the Bundesliga, indicating these countries as promising areas for targeted scouting.

Further analysis was carried out to determine which source leagues produced the most successful incoming players. The Portuguese top division, Liga Portugal, was found to yield the highest average post-transfer playing time, with players from this league averaging 63.8% of minutes played (see Figure 5). This strong performance suggests that Liga Portugal develops players who are well-prepared and adaptable to the demands of the Bundesliga, making it a valuable market for future signings. Transfers originating from within the Bundesliga also exhibited strong performance, averaging 57.4% of minutes played as shown in Figure 5. This consistency underscores the reliability of domestic transfers and validates ongoing intra-league recruitment strategies.

The analysis also extended to outbound loans, a critical mechanism for the development of players who are not yet part of the regular first-team setup. It was determined that players loaned from Germany gained the most playing time when placed in Spain's LaLiga and Belgium's Jupiler Pro League, with average playing percentages of 69.0% and 67.7%, respectively (see Figure 6). These findings highlight these leagues as particularly effective environments for providing meaningful match experience to develop players.

Collectively, the results from the EDA offer empirically grounded recommendations to

support the scouting and player development efforts at 1. FC Köln. The derived insights also contributed to enhancing the interpretability of the model used in subsequent phases of the project by contextualizing its performance and behavioral patterns.

3.2 Estimation Results

For the final model, an RMSE of 25.89 and a R^2 of around 0.276 was reached. This means the model is not the strongest predictor, but the results are not negligible either. In Figure 7, 50 examples of predictions can be seen compared to the actual playing percentage values. In a behavior-influenced domain like football, it is generally hard to have large predictive power, since outcomes are affected by many unobservable, or unknown factors. There are many hidden variables to consider, such as injuries post transfer, manager preferences/tactical fit, player attitude and disciplinary actions and team dynamics/chemistry. These all can have a detrimental effect on playing time but are hard to quantify well.

Regarding the approach made, predicting playing percentage for only the first year after the transfer also adds some noise, since two similar quality players may adapt at different speeds. Data quality is another important factor that plays a role in predictive power. The data used from *transfermarkt.de* had a significant number of missing values, and for some variables, variance is present in the way they were recorded for different players (height, market value).

Using SHapley Additive exPlanations (SHAP), the following interpretations were concluded in Figure 8 regarding the variables used in the model in order of feature importancies:

The age of the player during the transfer is the most impactful feature, where older players are predicted to play more than their younger counterparts as seen in Figure 9. The target country of the transfer also had meaningful differences between predictions, indicating that even when all other variables remain unchanged, where the player transfers to plays a huge role in determining their success. This remains in line with 1. FC Köln's hypothesis stating that players transferring from different leagues different success, providing further support to these assumptions. The same applies for the main position variable as well.

Interestingly, players transferring to higher market value teams are predicted to play less as shown on Figure 10. Possible explanations of this could be that higher market value teams play more games in a season, therefore they have bigger squad depth, and need to rotate their players more, and that they generally have higher quality players. The playing percentage from the season before the transfer also has an effect as expected. However, it is interesting to note that the relationship with the target variable is not linear. There is a significantly large difference between players who played 95 percent

plus minutes, compared to ones that played around 85 to 90 percent.

Players on loan are expected to play more, while players coming back from loan are expected to play less. This is fundamentally attributed to the fact that a lot of loan deals are made to help a player's development, and give them put them in a possibly weaker league, allowing them to gain more playing time and experience.

3.3 Dashboard

An interactive dashboard was developed to support scouting decisions by predicting the expected playing time of players in the season following a transfer, as shown in Figure 11. The tool is designed for ease of use and practical relevance in recruitment workflows. It was implemented using the Python framework Streamlit and connected to a GitHub repository to enable version control and continuous integration. Through deployment on Streamlit Cloud, the dashboard remains continuously available online[1]. For this, a Generalized Additive Model (GAM) was used with the predictions from XGBoost to capture any bias the model still has and to make the distribution less concentrated. This allows the dashboard to show larger differences between playing percentage, which helps in deciding between two players who would otherwise be very close in predicted success.

Users begin by entering the player's profile, including position, age, and market value. Transfer details are added by adjusting team market values to reflect the selling and buying clubs. Performance data from the previous season is then included, such as minutes played and combined goal contributions. The model outputs a predicted percentage of minutes played, which is then mapped to one of five role categories defined in collaboration with data analysts from 1. FC Köln to support practical interpretation in a scouting context:

- Below 20%: Not expected to play
- 20–39%: Expected to be a substitute
- 40–59%: Expected to be a rotation player
- 60–89%: Expected to be a key player
- 90% and above: Next star player

In the example of Florian Wirtz (Figure 11), the model predicts a playing time of approximately 68%, placing him in the "Expected to be a Key Player" category.

To support interpretation, the dashboard also displays the top three most similar historical transfers (Figure 12), filtered by position and competition level. All three are Bundesliga departures. Among them is Dominik Szoboszlai, now at Liverpool, who serves as a direct competitor to Wirtz in both role and context.

[1]Dashboard available at: 1. FC Köln Dashboard

4 Discussion

While recent research has focused on predicting transfer fees, modeling team-level dynamics, or analyzing network structures within the football transfer market (Dieles, Mattsson, & Takes, 2024; Dinsdale & Gallagher, 2023; Bosman, 2023), little attention has been paid to forecasting individual post-transfer playing time as a continuous outcome. This project addresses that gap by focusing on player-level performance, specifically, the percentage of minutes played in the first season after a transfer. The approach is distinct in that it models success not as a binary outcome but as a continuous metric, grounded in real-world scouting needs and supported by an interpretable prediction interface.

To support both modeling and scouting, an exploratory data analysis was conducted on transfer trends into the Bundesliga and outbound loans. It revealed that players from countries like Japan and Switzerland, as well as leagues such as Liga Portugal, tended to perform well post-transfer. Loan destination patterns also identified environments offering more playing time. These findings informed both the feature design for modeling and strategic scouting decisions.

The initial modeling approach used logistic regression to classify success as achieving more than 50 percent of possible minutes played. While this approach offered some insight, it oversimplified a complex problem. Playing time is a continuous and context-dependent outcome, and binary classification discarded useful information. The scope was therefore adjusted to predict percentage played as a continuous variable. Histogram-based gradient boosting was introduced to model non-linear relationships efficiently and to stabilize feature behavior across skewed distributions. This model provided a solid foundation but lacked the control and regularization needed to handle the full complexity of the dataset. The modeling approach was then extended to XGBoost, which is more effective at managing missing values, high-cardinality features, and deeper variable interactions. Hyperparameters were optimized using Bayesian search. The final model achieved an RMSE of 25.8 and an R^2 of 0.28, which is moderately high given the many unpredictable factors that influence player usage.

4.1 Limitations and Future Work

While the developed model provides valuable insights into the prediction of transfer success, several limitations remain. The primary data source, *transfermarkt.de*, is community-based and may be affected by potential inconsistencies in player valuations and gaps in historical data coverage (Coates & Parshakov, 2022). Although commonly used in academic studies, it lacks the granularity of proprietary data providers such as Opta, which offer detailed event-based match data (e.g., passes under pressure, progressive carries) (Stats Perform, 2024). In addition, the model does not incorporate a history of injury or future risk of injury, factors that can significantly impact the availability and perfor-

mance of a player. Due to privacy constraints and limited accessibility, such information is rarely available in public datasets. In addition, the model does not account for club-specific contextual variables. It treats transfers as context-neutral, although clubs differ in tactical systems, positional needs, and strategic objectives. Incorporating internal team information or tactical fit assessments could potentially improve model performance.

Future work could focus on the integration of advanced performance indicators, psychological or behavioral characteristics (e.g., adaptability, leadership), and market-related variables such as agent involvement or contractual details. Furthermore, extending the model to include coaching styles or intra-squad competition may offer a more comprehensive understanding of transfer success.

4.2 Managerial Implications

The model serves as a practical tool for 1. FC Köln to make more informed decisions around player transfers. By predicting the percentage of playing time a player is expected to receive after joining the club, and classifying the player into an expected squad role, it helps the scouting department in recruitment, performance analysis, and the sporting management in assessing the likelihood of a successful on-pitch transition. The tool can be used to narrow down the potential market to fewer players, who are a better fit for the club. This supports more strategic squad planning, helping to avoid situations where new signings "flop" (Ahmed, 2023).

From a financial perspective, making the right transfer can bring the club hundreds of thousands if not millions in transfer earnings and performance improvement, therefore it is crucial to use all the data available to make the best decisions.(Pantuso & Hvattum, 2019). While it will not replace scouting or coaching judgment, the model brings the objective outlook otherwise missing and provides a data-driven starting point. It can help decide loan decisions, tailor onboarding plans, and guide transfer negotiations, especially when assessing the risk of investing in players that the model does not rate highly (Chatziparaskevas, Bontempi, & Kotsiantis, 2024). Ultimately, it enables the club to manage both player value and performance outcomes more effectively, contributing to long-term sustainability and success.

4.3 Conclusion

This project demonstrates that predictive analytics can meaningfully support decision-making in professional football, supporting an affirmative answer to the research question. By applying machine learning techniques to large, historical datasets of football transfers, clubs like 1.FC Köln are provided with data-driven mechanisms to predict the success of incoming transfers. The results show that while precise predictions remain challenging due to behavioral and contextual variability, certain trends are robust: specific players

from specific nationalities and leagues tend to outperform others, and loan destinations significantly influence player development.

Scouting departments can now incorporate objective, model-backed insights into their evaluations, enhancing both the efficiency and accuracy of transfer decisions. Beyond the current implementation, the study lays the groundwork for future innovations in the data analytics department of 1. FC Köln, including the integration of more granular performance data, tactical fit measures, and psychological profiling. In a transcending era of football where data analysis plays a growing role not only regarding player transfers but also real-time match startegy, injury prevention and beyond, this project represents a strategic step of 1.FC Köln toward following the examples set by clubs such as *Brighton & Hove Albion* and *FC Liverpool* to win on and off the pitch in a competitive market (Ahmed, 2023).

A Appendix

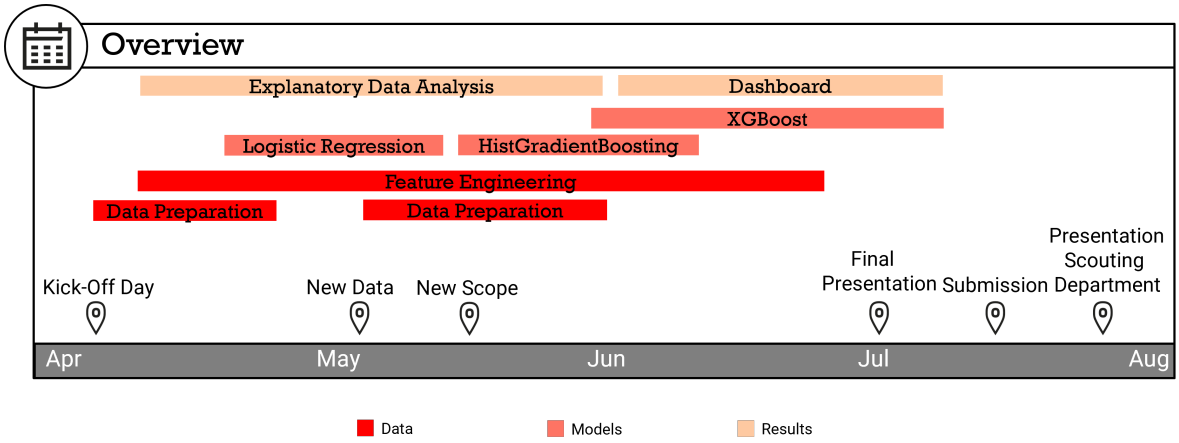

Figure 2: Timeline of the project

Table 1: Definition of the variables used for EDA and modeling

Variable Name	Description	Example Values
height	Height of the player in centimeters	178, 185, 192
transferAge	Age at time of transfer	21, 25, 29
nationality	Nationality of the players	German, Italian, French
percentage_played_before	Share of minutes played before transfer	25, 67, 95
percentage_played	Share of minutes played after transfer	0, 52, 89
success_before	Played 50% before transfer (1=yes)	0, 1
success	Played 50% after transfer (1=yes)	0, 1
isLoan	Was the transfer a loan?	0, 1
wasLoan	Previously on loan?	0, 1
was_joker	Often substituted in?	0, 1
foreign_transfer	Transfer involved foreign club?	0, 1
from_competition_competition_level	League level of previous club	1, 2, 3
to_competition_competition_level	League level of destination club	1, 2
foot	Preferred foot	left, right, both
mainPosition	Primary position	centre-back, right-wing
positionGroup	Position group	defender, midfielder, attacker
from_competition_competition_area	Region of previous club	Germany, France, Other
to_competition_competition_area	Region of new club	Germany, Netherlands, Other
scorer_before_grouped_category	Grouped scorer points without defenders and goalkeepers	0–3, 4–6, 7–10
clean_sheets_before_grouped	Grouped clean sheets	0–1, 2–4, 5–9
fromTeam_marketValue	Market value of old team [€]	210M, 450M, 780M
toTeam_marketValue	Market value of new team [€]	180M, 600M, 1100M
marketvalue_closest	Player value at transfer [€]	5M, 18M, 35M
value_per_age	Market Value ÷ age [€]	0.3M, 0.7M, 1.2M
value_age_product	Market Value × age [€]	120M, 240M, 350M
team_market_value_relation	Destination/origin value ratio	–1.2, 0.0, 1.3

Table 2: Logistic Regression Performance by Class

Class	Precision	Recall	F1-Score
0 (no success)	0.66	0.47	0.55
1 (success)	0.70	0.83	0.76
Accuracy		0.69	

Table 3: Best Hyperparameters Found by Bayesian Optimization

Hyperparameter	**Value**
`colsample_bytree`	0.4820
`gamma`	2.0000
`learning_rate`	0.0190
`max_depth`	6
`min_child_weight`	17
`n_estimators`	1083
`reg_alpha`	2.0000
`reg_lambda`	1.6093
`subsample`	0.8457

Table 4: HistGradientBoostingClassifier Performance by Class

Class	**Precision**	**Recall**	**F1-Score**
0 (no success)	0.74	0.68	0.71
1 (success)	0.61	0.67	0.64
Accuracy	0.68		

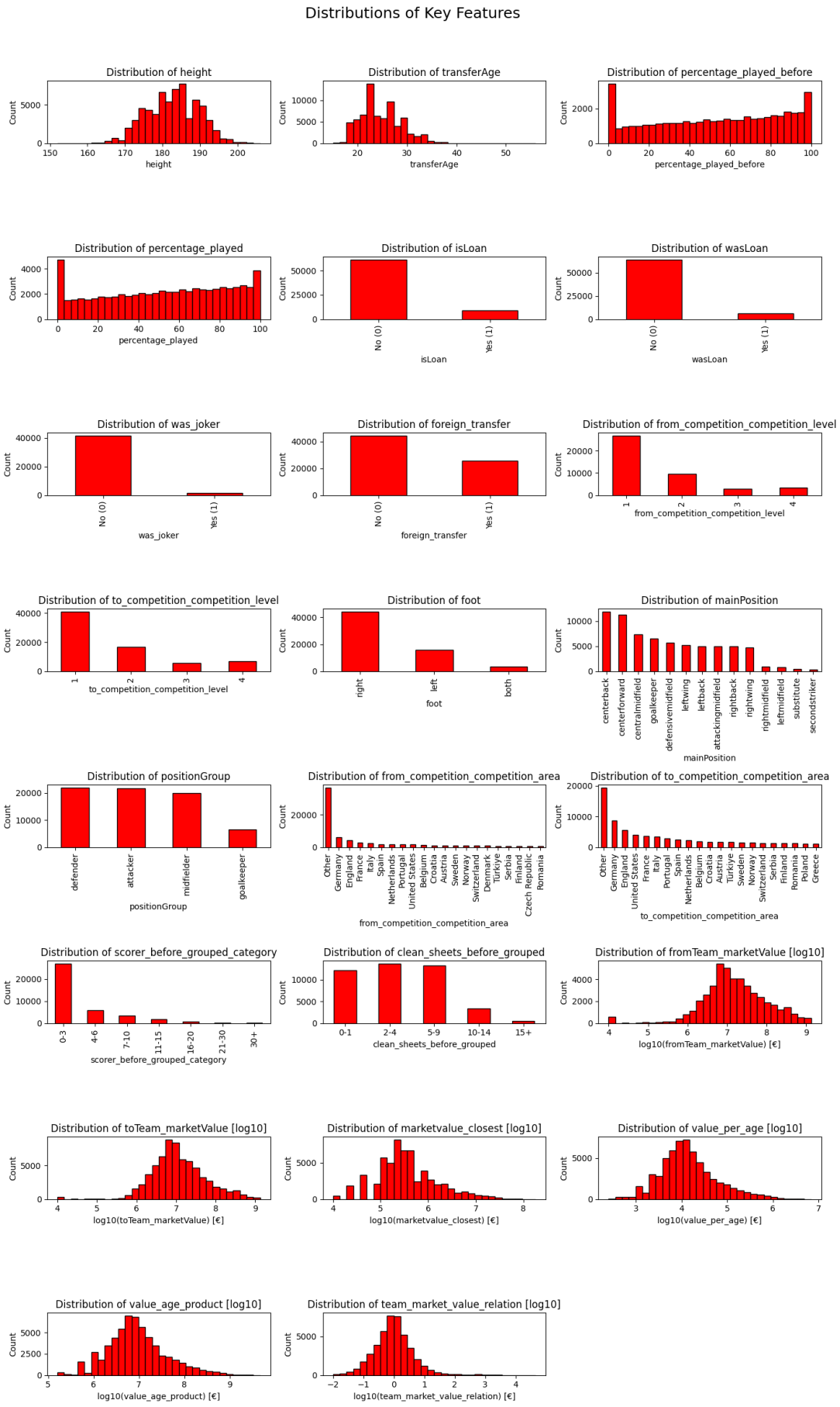

Figure 3: Distribution of Model Features

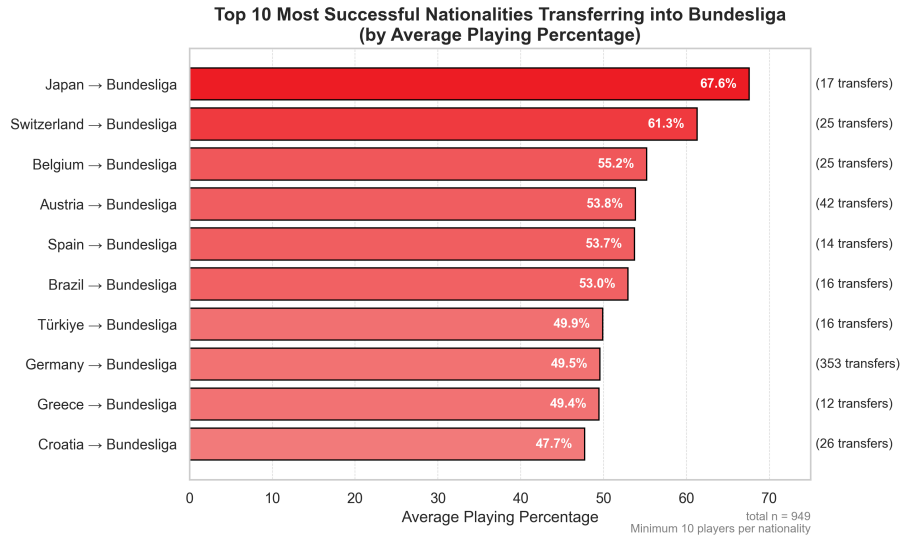

Figure 4: Top 10 Most Successful Nationalities Transferring into Bundesliga.

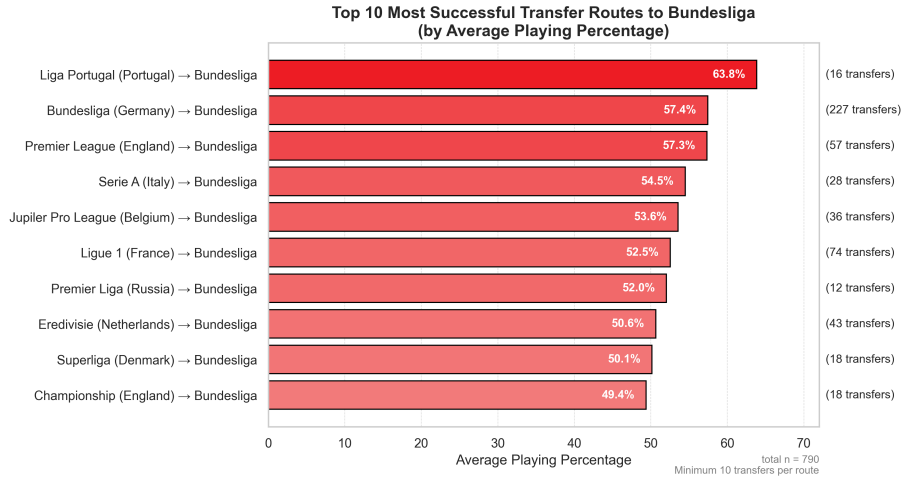

Figure 5: Top 10 Most Successful Transfer Routes to Bundesliga.

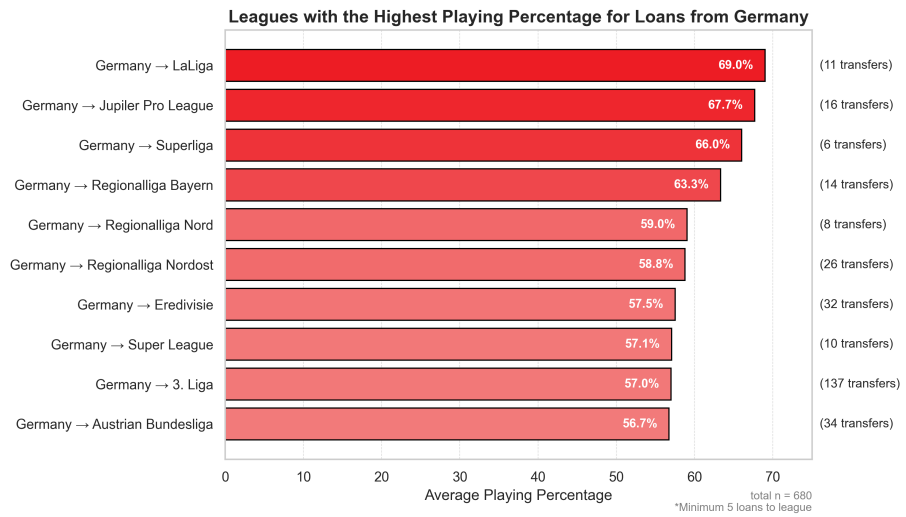

Figure 6: Leagues with the Highest Playing Percentage for Loans from Germany

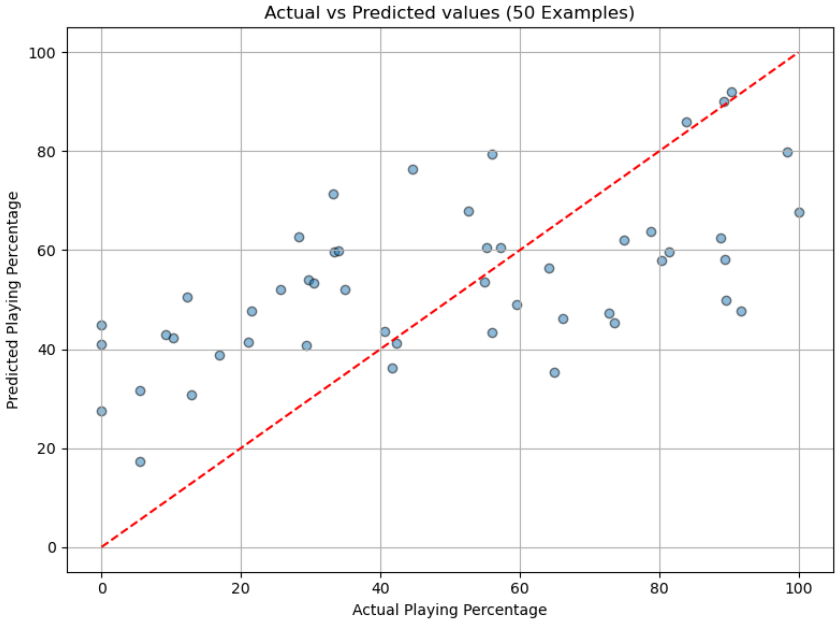

Figure 7: Actual vs. predicted playing percentage values using the final XGBoost model. The diagonal line represents perfect predictions.

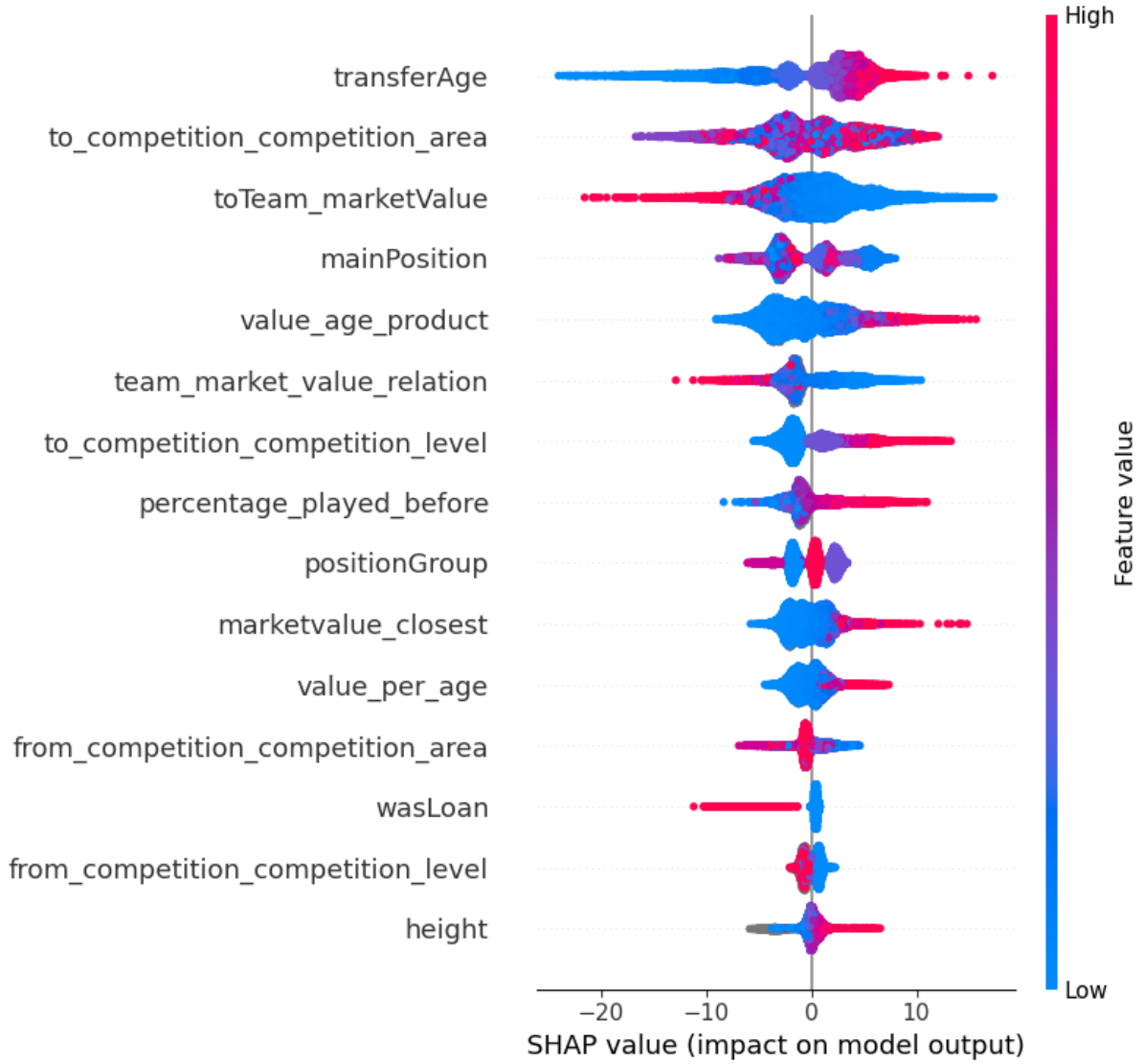

Figure 8: SHAP values for the 15 most important variables

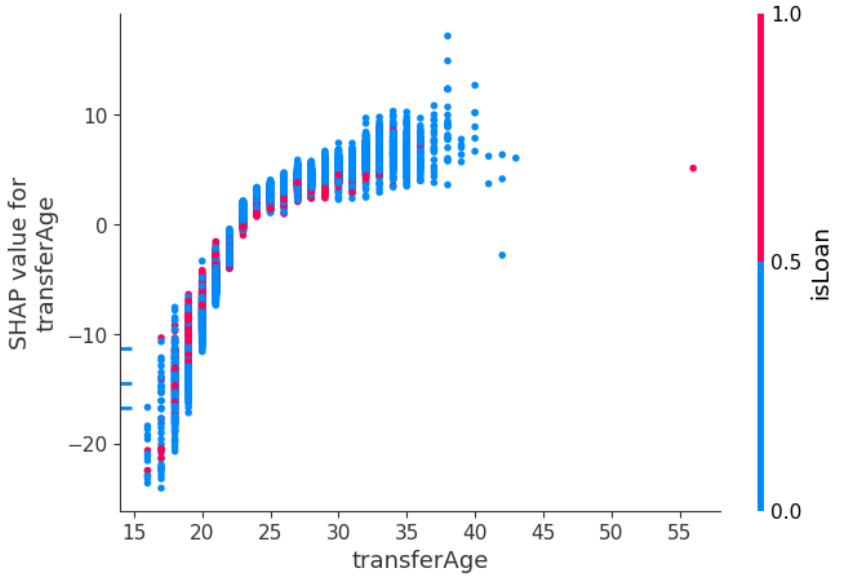

Figure 9: SHAP dependence plot for transfer age. It shows that older players tend to have higher predicted playing percentages post-transfer Loan player predictions are displayed in red. It is important to note that for loan players age has a slightly different effect on predictions.

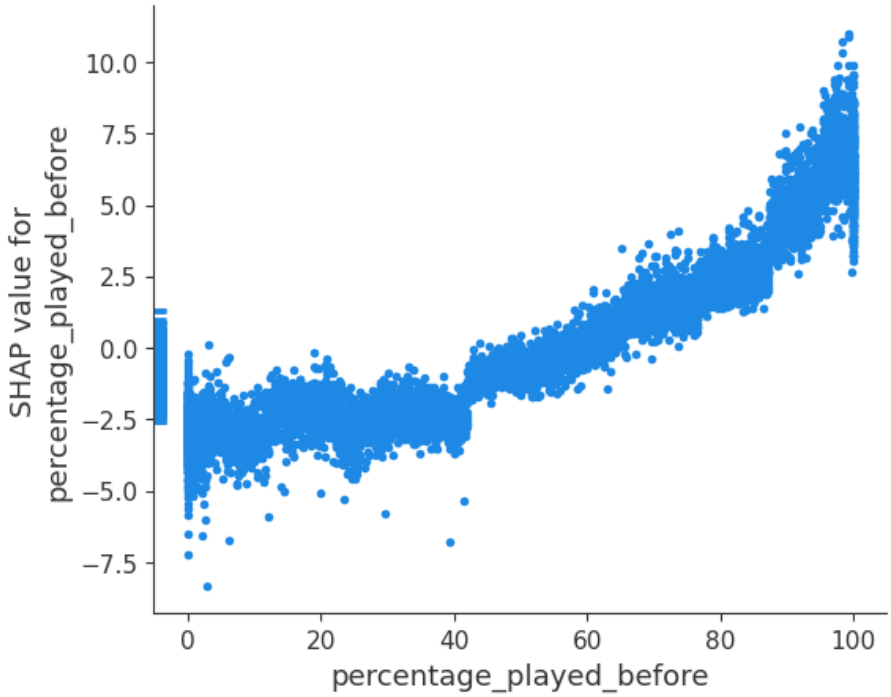

Figure 10: SHAP summary plot showing how previous predicted playing percentage affects the playing percentage after the transfer.

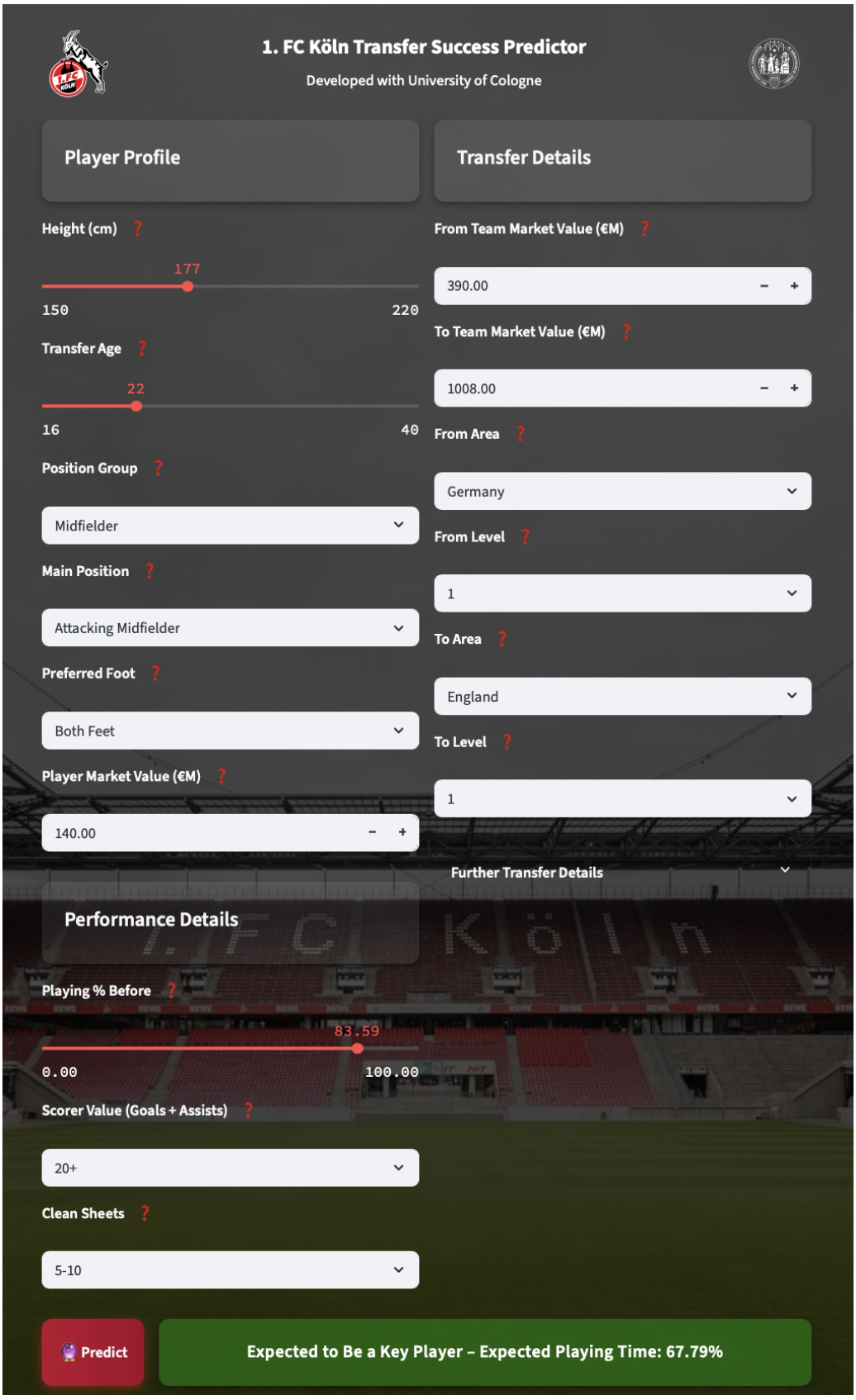

Figure 11: Interactive Dashboard Used for Predicting Player Transfer Outcomes

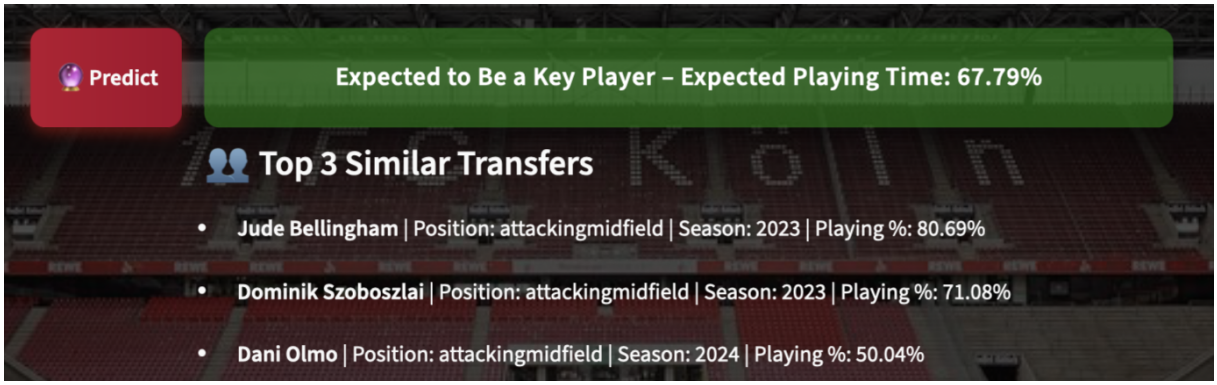

Figure 12: Similar Players for Florian Wirtz

References

1. FC Köln. (2024a). *Fc hat 140.000 mitglieder.* Retrieved from `https://fc.de/aktuelles/news/fc-hat-140-000-mitglieder` (Accessed: 2025-07-11)

1. FC Köln. (2024b). *Geschäftsbericht 2023/24.* Retrieved from `https://fc.de` (Accessed: 2025-07-11)

Ahmed, M. (2023). *Liverpool's secret weapon: the data analyst who helped jürgen klopp make history.* Financial Times. Retrieved from `https://www.ft.com/content/c8b3bba3-6051-43eb-b176-6cefa4d9b9f5` (Accessed: 2025-07-13)

Barajas, , & Rodriguez, P. (2010). Spanish football clubs' finances: Crisis and player salaries. *International Journal of Sport Finance*, *5*, 52-66.

Bosman, D. (2023). *Predicting fees in soccer european leagues: A machine and deep learning transfer model.*

Chai, T., & Draxler, R. R. (2014). Root mean square error (rmse) or mean absolute error (mae)?–arguments against avoiding rmse in the literature. *Geoscientific Model Development*, *7*(3), 1247–1250. doi: 10.5194/gmd-7-1247-2014

Chatziparaskevas, N., Bontempi, G., & Kotsiantis, S. (2024). The impact of information systems and data science on player recruitment in professional football. In *Aip conference proceedings* (Vol. 3220, p. 050011). AIP Publishing. Retrieved from `https://pubs.aip.org/aip/acp/article/3220/1/050011/3315890` doi: 10.1063/5.0197106

Chen, T., & contributors. (2025). Xgboost documentation [Computer software manual]. Retrieved from `https://xgboost.readthedocs.io/en/stable/` (Accessed: 2025-07-13)

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785–794). doi: 10.1145/2939672.2939785

Coates, D., & Parshakov, P. (2022). Valuation of football players using crowdsourced transfer fees. *International Journal of Sport Finance*, *17*(1), 1–16. Retrieved from `https://hcapps.holycross.edu/hcs/RePEc/fek/Session08.2-Coates.pdf` (Accessed: 2025-07-12)

Dieles, T., Mattsson, C., & Takes, F. (2024). Identifying successful football teams in the european player transfer network. *Applied Network Science*, *9*(65). Retrieved from `https://doi.org/10.1007/s41109-024-00675-7` doi: 10.1007/s41109-024-00675-7

Dinsdale, D., & Gallagher, J. (2023). *Transfer portal: Accurately forecasting the impact of a player transfer in soccer.*

Head, T., MechCoder, et al. (2021). *Scikit-optimize: Sequential model-based optimization in python.* `https://scikit-optimize.github.io/`. (Accessed: 2025-07-13)

Jarjabka, A., Fűrész, D. I., & Havran, Z. (2024). The impact of cultural distance on the migration of professional athletes as high-skilled employees. *Economia e Politica Industriale / Journal of Industrial and Business Economics*, *51*(3), 585–603. Retrieved from `https://link.springer.com/article/10.1007/s40812-023-00288-8` doi: 10.1007/s40812-023-00288-8

Kicker. (2023, December). *Cas bestätigt transfersperre für den 1. fc köln.* Retrieved from `https://www.kicker.de/cas-bestaetigt-transfersperre-fuer-den-1-fc-koeln-986297/artikel` (Accessed: 2025-07-11)

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in neural information processing systems* (Vol. 30). Retrieved from `https://proceedings.neurips.cc/paper_files/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html`

Pantuso, G., & Hvattum, L. M. (2019). A chance-constrained model for football team management. *arXiv preprint arXiv:1911.04689*. Retrieved from `https://arxiv.org/abs/1911.04689`

Prockl, F., & Frick, B. (2018). Information precision in online communities: Player valuations on www.transfermarkt.de. *International Journal of Sport Finance*, *13*(4), 319–335.

Scikit-learn. (2024). *Histogram-based gradient boosting.* Retrieved from `https://scikit-learn.org/stable/modules/ensemble.html#histogram-based-gradient-boosting` (Accessed: 2025-07-11)

Snoek, J., Larochelle, H., & Adams, R. P. (2012). Practical bayesian optimization of machine learning algorithms. *arXiv preprint arXiv:1206.2944*. Retrieved from `https://doi.org/10.48550/arXiv.1206.2944` doi: 10.48550/arXiv.1206.2944

Stats Perform. (2024). *Opta analytics — performance insights for football.* Retrieved from `https://www.statsperform.com/opta-analytics/` (Accessed: 2025-07-11)

Szymanski, S. (2010). The market for soccer players in england after bosman: Winners and losers. In *The comparative economics of sport* (pp. 17–18). Palgrave Macmillan.