Sports Analytics: Predicting Player Value and Injury Risk

Vaishnavi Kamdi
George Washington University

Professor David W. Trott

Abstract

This project focuses on applying classical machine learning techniques to address two critical aspects of sports analytics: predicting football (soccer) player market value through regression and assessing player injury risk via classification. By analyzing datasets incorporating player attributes, historical performance, market valuations, and injury records, the study aims to develop robust predictive models. For player value prediction, regression models including Linear Regression, Random Forest, and XGBoost are explored, with a focus on features like age, physical attributes, and historical peak market value. For injury risk assessment, classification models such as Logistic Regression, Random Forest, and XGBoost are employed, utilizing features like player age, BMI, and historical injury data. The project emphasizes thorough data preprocessing, feature engineering to derive insightful player metrics, and rigorous model evaluation using MAE, RMSE, and R^2 for regression, and AUC-ROC, F1-score, and precision-recall metrics for classification. The ultimate goal is to provide a comparative analysis of model performance and deliver a Streamlit dashboard for interactive player value and injury risk predictions, thereby aiding in data-driven decision-making for player acquisitions and health management.

Keywords: Sports Analytics, Machine Learning, Player Valuation, Injury Prediction, Regression, Classification, Football Analytics, Transfermarkt, FIFA

Contents

1 Introduction

The domain of sports analytics has seen a significant surge with the advent of advanced machine learning techniques, enabling deeper insights into player performance, valuation, and well-being. This project aims to contribute to this field by developing predictive models for two critical metrics in football (soccer): player market value and injury risk. Accurately estimating a player's market value is crucial for clubs during transfer negotiations and for financial planning [7].

Simultaneously, predicting the likelihood of a player sustaining an injury can inform training regimens, player rotation strategies, and overall health management, ultimately impacting team performance and player career longevity [14].

This study leverages historical player data, including performance statistics, physical attributes, injury records, and market valuation trends, sourced from publicly available datasets such as Transfermarkt and FIFA. The core of this project involves two primary predictive tasks:

1. **Player Market Value Prediction:** A regression task to estimate a player's transfer market value. This involves exploring models like Linear Regression, Random Forest Regressor, and XGBoost Regressor.

2. **Injury Risk Assessment:** A classification task to predict the likelihood of a player sustaining a significant injury. This will be approached using Logistic Regression, Random Forest Classifier, and XGBoost Classifier.

The project emphasizes a comprehensive methodology encompassing thorough data preprocessing, exploratory data analysis (EDA) to uncover underlying patterns, and strategic feature engineering to create new, informative variables. Model performance will be rigorously evaluated using appropriate metrics for each task, such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R^2 for regression, and Area Under the ROC Curve (AUC-ROC), F1-score, and precision-recall curves for classification. Furthermore, model interpretability will be explored through feature importance analysis.

The expected deliverables include a comparative analysis of the performance of different algorithms for both prediction tasks and an interactive Streamlit dashboard designed to provide predictions for player value and injury risk based on user inputs. This work seeks to provide actionable, data-driven insights for football clubs, analysts, and enthusiasts.

2 Background and Motivation

2.1 Background

The professional football industry operates as a multi-billion dollar global market where player assets are paramount [9]. Player valuation is a complex process influenced by a multitude of factors including on-field performance, age, contract duration, marketability, and injury history [5, 7]. Traditional scouting methods, while valuable, are often subjective and can be augmented by data-driven approaches to provide more objective assessments. Machine learning offers the potential to model these complex relationships and provide quantitative estimates of player market value [1].

Player injuries represent another significant challenge in professional sports [14]. Injuries not only impact a player's availability and performance but also impose substantial financial burdens on clubs through medical expenses and the potential depreciation of player value. Understanding the factors contributing to injury risk and developing predictive models can help clubs implement preventative measures, optimize training loads, and make informed decisions regarding player recruitment and squad management [4, 3].

2.2 Motivation

The motivation for this project stems from the increasing need for football clubs and associated stakeholders to make more informed, data-driven decisions.

- **For Player Valuation:** Clubs require accurate valuation models to navigate the transfer market effectively, identifying undervalued talents or determining fair prices for their own players [17]. Agents and players can also benefit from objective valuation benchmarks during contract negotiations.

- **For Injury Risk Assessment:** Proactive injury prevention is key to maintaining squad strength and player welfare. Predictive models can serve as early warning systems, allowing medical staff and coaches to intervene with personalized training or rest protocols, potentially reducing injury incidence and severity [3, 13].

This project aims to address these needs by applying robust machine learning methodologies to real-world football data, providing tools that can enhance strategic decision-making in both financial and player management aspects of the sport. The development of an interactive dashboard further aims to make these analytical insights accessible to a broader audience.

2.3 Proposed Methodology/Solution

The project will tackle two distinct predictive tasks:

1. **Player Market Value Prediction (Regression):**
   - Utilize player attributes (age, position, height, weight), performance metrics, and historical market value data from Transfermarkt.
   - Implement and compare Linear Regression, Random Forest Regressor, and XGBoost Regressor.
   - Focus on feature engineering to capture nuances like career stage (derived from age) and value trajectory.

2. **Player Injury Risk Prediction (Classification):**
   - Employ data from the FIFA dataset, including player physical attributes, workload indicators (minutes played, games played), and historical injury data.
   - The target variable will be a binary indicator of whether a player experienced a "significant injury" in the previous season.
   - Implement and compare Logistic Regression, Random Forest Classifier, and XGBoost Classifier.
   - Feature engineering will include calculating metrics like Body Mass Index (BMI), injury rates, and cumulative workload.

A key component of the methodology will be rigorous data preprocessing, including handling missing values, encoding categorical features, and feature scaling. Model evaluation will use standard metrics (MAE, RMSE, R^2 for regression; AUC-ROC, F1-score, Precision, Recall for classification). Feature importance analysis will be conducted to understand the key drivers behind the predictions. Finally, a Streamlit application will be developed to showcase the models' predictive capabilities.

2.4 Related Works

The application of machine learning in sports analytics, particularly football, has gained considerable traction. Several studies have explored player valuation using diverse modeling techniques. Early approaches often involved econometric or linear models to identify key value drivers [5, 6]. For instance, Müller et al. (2017) [1] demonstrated the use of machine learning techniques, incorporating factors like age, on-field performance metrics (goals, passes), and duels won, to move beyond subjective crowd judgments in player valuation. Similarly, Al Asadi and Taşdemir (2023) [2] explored various regression models, including multiple linear regression and ensemble methods, to estimate player market values using comprehensive datasets that often include player statistics, age, and contract status [7]. Liu et al. (2024) [16] compared Random Forest, Gradient Boosting, and Ridge Regression, finding non-linear models often superior for this task. Other approaches have utilized K-Nearest Neighbors (KNN) based on player similarity for wage or value estimation [8]. Research has also focused on predicting future market value trajectories using AI techniques [9]. These studies highlight the capacity of data-driven models to capture the complex determinants of a player's economic worth. This project aims to build upon such approaches by leveraging publicly available data and a broad range of engineered features.

In the realm of injury prediction, research has increasingly focused on identifying risk factors and employing predictive modeling. Training load, often monitored via GPS, is a commonly investigated factor [4]. Lopes et al. (2024) [3] introduced a machine learning approach using GPS data alongside player-specific parameters (like position, session type) to predict noncontact injuries. Other studies have utilized different data sources, such as preseason screening tests combined with models like XGBoost [10], or focused on predicting specific injury types like hamstring strains [11]. Some research explores non-technical data collection methods [12] or employs models like logistic regression to identify risk factors such as age or kinesiophobia [13]. These studies often point to the multifactorial nature of injuries [14].

A common thread in recent literature, including systematic reviews [15], is the move towards more sophisticated models like gradient boosting (XGBoost) and ensemble methods (Random Forest) for both valuation and injury prediction, confirming the relevance of the models chosen for this project. Furthermore, the emphasis on creating an interactive tool for practical application distinguishes this project from purely academic explorations.

3 Data Pre-processing and Feature Engineering

3.1 Data Sources

This project utilizes two primary datasets sourced from publicly available platforms, providing a rich collection of player attributes, market information, and injury records.

3.1.1 Transfermarkt Player Dataset

- **Source:** Kaggle Dataset (https://www.kaggle.com/datasets/davidcariboo/player-scores). This dataset contains player market value information originally sourced or compiled from Transfermarkt (www.transfermarkt.com).
- **Content:** This dataset contains information on players, including their `player_id`, name, `current_club_id`, `last_season` played, `height_in_cm`, `foot`, `position`, date of birth (from which `age` is derived), and historical `market_value_in_eur` and `highest_market_value_in_eur`.
- **Purpose:** Primarily used for the player market value prediction task (regression).
- **Dataset Used:** The specific dataset utilized for the modeling phase was the `cleaned_tm_players_dataset_v3_with_features.csv` file, which represents the data after the cleaning and feature engineering steps described in the subsequent sections were applied.

3.1.2 FIFA Injury Dataset

- **Source:** Kaggle Dataset (https://www.kaggle.com/datasets/stefanoleone992/fifa-23-complete-player-dataset). This dataset contains player attributes, ratings, and statistics, likely derived from the FIFA video game series. Injury data was either included or engineered from related metrics.
- **Content:** This dataset includes player attributes such as `age`, `height_cm`, `weight_kg`, `position`, FIFA ratings (`fifa_rating`, `pace`, `physic`), playing statistics (`season_minutes_played`, `season_games_played`, `season_matches_in_squad`), and detailed injury history (`season_days_injured`, `total_days_injured`, and engineered features related to previous season injuries).
- **Purpose:** Primarily used for the player injury risk assessment task (classification).
- **Dataset Used:** The modeling for injury risk utilized the `cleaned_fifa_dataset_v3_with_features.csv` file, which incorporates the results of the data cleaning and feature engineering processes detailed later in this section.

3.2 Exploratory Data Analysis (EDA)

EDA was performed on both datasets to understand data distributions, identify missing values, uncover relationships between variables, and guide feature engineering and modeling choices.

3.2.1 Transfermarkt Player Dataset EDA

The Transfermarkt dataset was analyzed to understand player demographics and market value distributions.

- **Missing Values:** An initial assessment of missing values in the raw Transfermarkt dataset revealed that several columns, such as `current_club_id`, `height_in_cm`, `foot`, and `agent_name`, had missing entries. These were addressed during the data cleaning phase through imputation or removal based on the extent of missingness and feature importance. Figure 1 illustrates the missing data patterns.

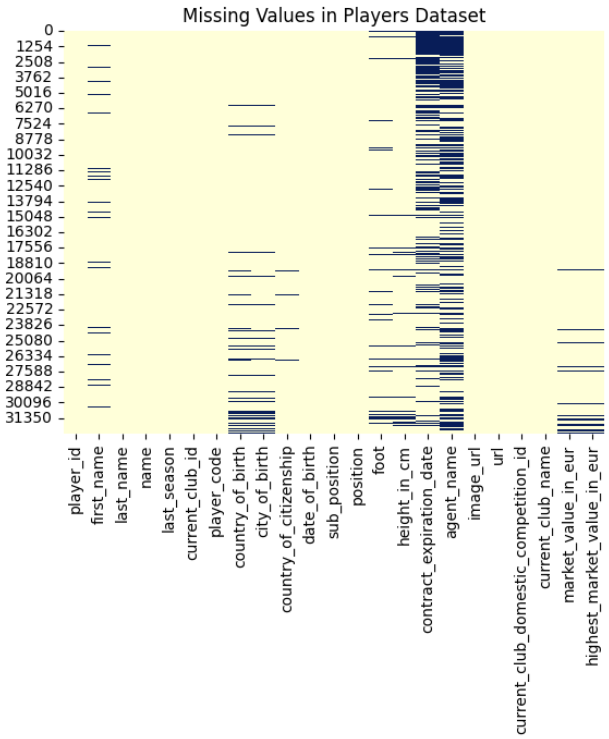

Figure 1: Missing Value Heatmap for the initial Transfermarkt Players Dataset, showing patterns of missing data across various features before cleaning.

- **Distribution of Player Ages:** Player ages typically follow a distribution skewed towards younger players, with a peak in the mid-twenties, reflecting the general career trajectory in professional football.

- **Distribution of Player Heights:** Player heights are generally normally distributed, centered around the average height for football players, though variations exist by position.

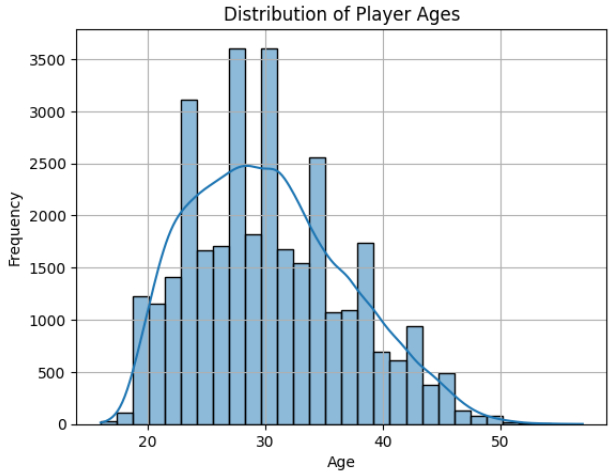

Figure 2: Distribution of Player Ages in the Transfermarkt Dataset. This histogram shows a peak age range typically between 25-30 years, with a decline in older age groups.

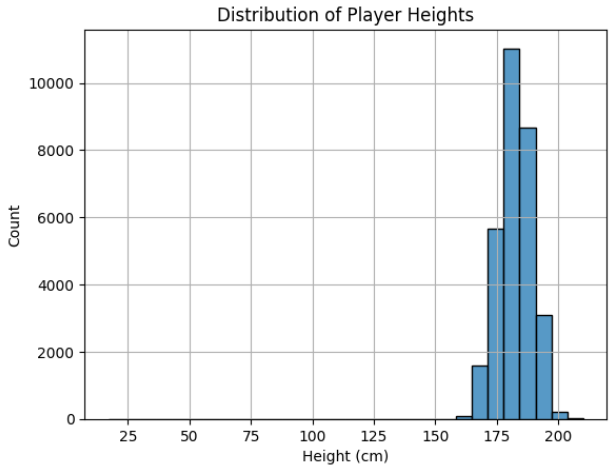

Figure 3: Distribution of Player Heights in the Transfermarkt Dataset. Most players fall within the 175-190 cm range.

- **Player Positions:** The dataset contains a variety of player positions, with midfielders and defenders often being the most numerous.

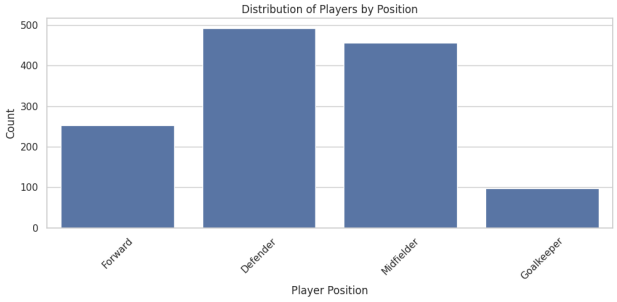

Figure 4: Distribution of Player Positions in the Transfermarkt Dataset. Defenders and Midfielders are the most common positions, followed by Forwards and Goalkeepers.

- **Market Value Distribution:** The target variable, `market_value_in_eur`, is typically right-skewed, with a few players having exceptionally high market values. This skewness often necessitates transformations (like log transformation) for regression modeling.

- **Market Value vs. Age:** The relationship between market value and age is non-linear, generally increasing until a player reaches their peak years (mid to late twenties) and then declining.

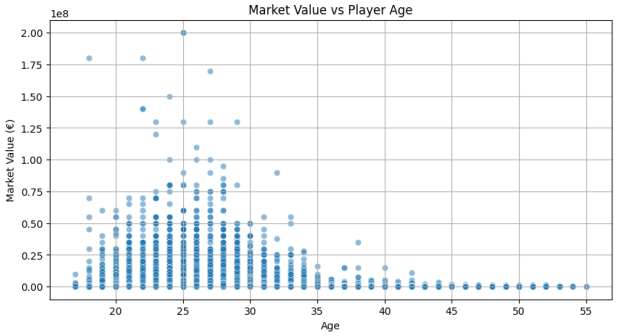

Figure 5: Scatter plot of Market Value against Player Age in the Transfermarkt Dataset, illustrating the typical career value curve.

- **Market Value by Position:** Market values can vary significantly by player position, with attacking players often commanding higher values.

- **Correlations:** A correlation heatmap revealed relationships between numerical features. `highest_market_value_in_eur` showed a strong positive correlation with the current `market_value_in_eur`. `Age` showed a more complex, often non-linear relationship with market value, typically peaking in the late twenties.

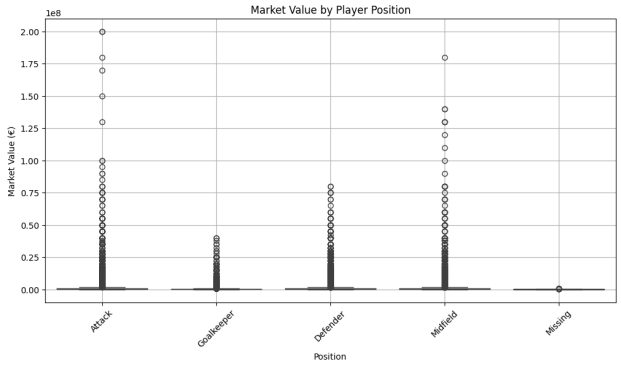

Figure 6: Market Value by Player Position in the Transfermarkt Dataset. Attackers and Midfielders show a wider range and higher median market values compared to Defenders and Goalkeepers. Outliers with very high market values are visible across attacking and midfield positions.

Figure 7: Correlation Matrix of Numeric Features in the Transfermarkt Dataset. This heatmap shows a strong positive correlation (0.78) between `market_value_in_eur` and `highest_market_value_in_eur`. `Age` has a negative correlation with `player_id` and `last_season` (expected as IDs/seasons progress) and a slight negative correlation with current market value in this raw view, though its true impact is often curvilinear.

3.2.2 FIFA Injury Dataset EDA

The FIFA dataset was explored to understand player characteristics, workload, and injury patterns.

- **Missing Values:** The dataset contained missing values in several columns, particularly for detailed performance stats or historical injury data that might not be available for all players or seasons. The `missingno` library was used to visualize this.

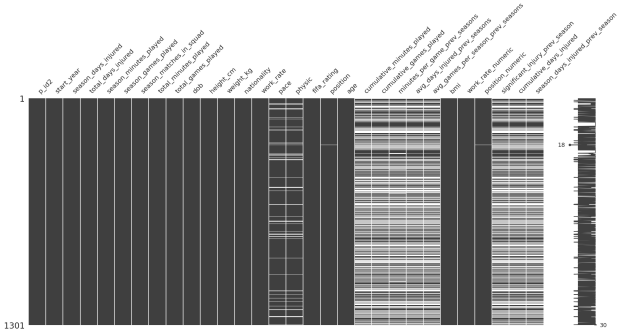

Figure 8: Missing Value Matrix for the FIFA Injury Dataset. This plot highlights columns with significant missing data, such as `pace`, `physic`, and several cumulative/previous season metrics, which required careful handling (imputation or removal).

- **Distribution of FIFA Ratings:** Player ratings (e.g., `fifa_rating`) are generally normally distributed, centered around an average skill level.

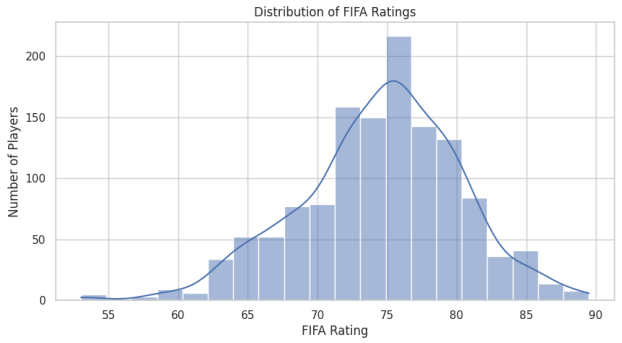

Figure 9: Distribution of FIFA Overall Ratings. Ratings are somewhat normally distributed, with a peak around 75-80.

- **FIFA Rating by Position (Outlier View):** A boxenplot provides a more detailed view of the distribution of FIFA ratings across different player positions, highlighting medians, quartiles, and potential outliers.

- **Pairwise Relationships of Key Attributes:** A pair plot was used to visualize the relationships and distributions between key attributes like `fifa_rating`, `pace`, `physic`, `age`, and `bmi`. This helps in identifying correlations and patterns at a glance. The diagonal shows the distribution of each attribute.

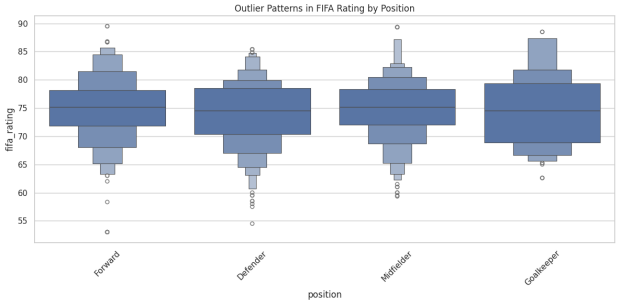

Figure 10: FIFA Rating by Player Position (Boxenplot). This view emphasizes the distribution shape and outliers for ratings within each position category.

- **Distribution of Age (FIFA Dataset):** Similar to the Transfermarkt dataset, player ages show a peak in the mid to late twenties.

- **Distribution of Days Injured:** The number of `season_days_injured` is often heavily skewed, with many players having zero or few days injured, and a long tail of players with many days injured.

- **Player Positions (FIFA Dataset):** The distribution of player positions was also examined, showing a similar pattern to the Transfermarkt dataset with Defenders and Midfielders being most frequent.

- **Target Variable: `significant_injury_prev_season`:** The balance of the target class for injury prediction was checked. This is important for understanding potential class imbalance issues.

- **Correlations (FIFA Dataset):** A correlation heatmap highlighted relationships between various player attributes, workload indicators, and injury metrics. For example, `season_minutes_played` and `season_games_played` are highly correlated. `Age` might show correlations with cumulative injury metrics.

This EDA phase was crucial for understanding the datasets' characteristics and informing the subsequent feature engineering, data cleaning, and modeling steps.

3.3 Feature Engineering

To enhance the predictive power of the models, several new features were derived from the existing data in both datasets.

3.3.1 Transfermarkt Player Dataset Features

For the player market value prediction task, the primary engineered feature was:

- **Age (`age`):** Calculated from the player's date of birth and the date of the market value record. Age is a critical factor in player valuation, often exhibiting a non-linear relationship with market value.

- **Position Encoded (`position_encoded`):** Player positions (e.g., Forward, Midfielder, Defender, Goalkeeper) were numerically encoded to be used in machine learning models.

3.3.2 FIFA Injury Dataset Features

For the injury risk classification task, the following features were engineered:

- **Body Mass Index (`bmi`):** Calculated as weight (kg) / (height (m)^2). BMI can be an indicator of physical condition.
- **BMI Class (`bmi_class`):** Players were categorized into 'Underweight', 'Normal', or 'Overweight' based on their BMI.
- **Injury Rate (`injury_rate`):** Calculated as the player's total days injured per game played, aiming to quantify injury susceptibility relative to exposure.
- **Minutes Per Game (`minutes_per_game`):** Reflects player workload intensity or efficiency, calculated by dividing total minutes played by games played in a season.
- **Experience Years (`experience_years`):** Estimated as the difference between the current season year and the player's year of birth, serving as a proxy for professional experience.
- **Cumulative Injury Metrics:** Features such as `cumulative_days_injured` and `cumulative_games_played` were created to capture a player's historical load and injury burden.
- **Previous Season Injury Metrics:** Features like `avg_days_injured_prev_seasons`, `avg_games_per_season_prev_seasons`, and `season_days_injured_prev_season` were engineered to incorporate recent injury history and playing patterns.
- **Work Rate Numeric (`work_rate_numeric`):** If player work rates (e.g., High/Medium/Low) were available, they were numerically encoded.
- **Position Numeric (`position_numeric`):** Player positions were numerically encoded.

These engineered features were designed to provide more nuanced information to the models than raw data alone.

3.4 Data Cleaning and Preparation

Both datasets underwent a series of cleaning and preparation steps to make them suitable for modeling.

- **Handling Missing Values:** Missing values were addressed based on the nature of the feature and the extent of missingness. Strategies included:
  - Dropping columns with a very high percentage of missing values if they were not deemed critical or could not be reliably imputed (e.g., `agent_name` in Transfermarkt).
  - Imputing missing numerical values using mean or median (e.g., for `height_in_cm` in Transfermarkt, or certain FIFA attributes after careful consideration).
  - Imputing missing categorical values using the mode or a placeholder category.
  - For some injury-related historical features in the FIFA dataset, missing values (NaNs) were filled with 0, assuming no prior record meant no injuries or games for that specific metric.
- **Data Type Conversion:** Columns were converted to appropriate data types (e.g., string dates to datetime objects for age calculation, numerical features to float or int).
- **Outlier Treatment:** While not explicitly detailed as a separate step in the notebooks for systematic removal, outlier detection was part of EDA (e.g., Figure 6, Figure 10). Extreme outliers in the target variable (market value) were a motivation for log transformation. For predictor variables, tree-based models (Random Forest, XGBoost) are generally robust to outliers.
- **Categorical Feature Encoding:** Categorical features like player position, foot, and BMI class were converted into numerical representations using techniques such as one-hot encoding or label encoding, depending on the feature's cardinality and the model requirements.
- **Duplicate Removal:** Duplicate rows were checked for and removed if present.
- **Filtering:** Irrelevant data or rows not meeting specific criteria (e.g., players with no market value data) were filtered out.

These steps ensured data quality and consistency, preparing the datasets for the subsequent modeling phase.

4 Methodology: Predictive Modeling

This project employs supervised machine learning techniques for two distinct tasks: predicting player market value (a regression problem) and assessing player injury risk (a classification problem).

4.1 Player Market Value Prediction (Regression)

The goal is to predict the `market_value_in_eur` for players using the Transfermarkt dataset.

4.1.1 Selected Features

Based on EDA and feature engineering, the primary features considered for the regression models included:

- `age`
- `height_in_cm`
- `position_encoded` (numerical representation of player position)
- `highest_market_value_in_eur` (player's historical peak market value)

The importance of factors like age and historical value aligns with findings in previous valuation studies [1, 5].

4.1.2 Models Implemented

The following regression algorithms were implemented and evaluated:

- **Linear Regression:** As a baseline model to capture linear relationships [5].
- **Random Forest Regressor:** An ensemble learning method known for its robustness and ability to handle non-linearities, commonly applied in sports analytics [15].
- **XGBoost Regressor:** A gradient boosting algorithm, often providing high performance due to its regularization and efficiency, frequently used for prediction tasks [2, 15].

4.1.3 Handling Data Leakage and Log Transformation

- **Data Leakage Mitigation:** Initial modeling revealed potential data leakage when a feature like `value_log` (log of current market value) was inadvertently included as a predictor for the current market value itself, leading to overly optimistic performance metrics. This feature was subsequently removed in retraining to ensure a realistic evaluation of model performance based on genuinely independent features.
- **Log Transformation of Target Variable:** The distribution of `market_value_in_eur` was observed to be highly right-skewed. To address this and stabilize variance, the natural logarithm of the market value (`value_log`) was used as the target variable in some modeling iterations. Predictions made on the log scale were then inverse-transformed (exponentiated) back to the original euro scale for evaluation. This technique is common when dealing with skewed financial data.

4.2 Player Injury Risk Prediction (Classification)

The objective is to classify whether a player is at risk of a significant injury using the FIFA dataset.

4.2.1 Target Variable Definition

The target variable for this task was `significant_injury_prev_season`, a binary feature indicating whether a player experienced a significant injury in the preceding season (1 for injury, 0 for no injury). This was derived or directly available in the `cleaned_fifa_dataset_v3_with_features.csv`.

4.2.2 Selected Features

Key features used for the classification models, drawn from the engineered FIFA dataset, included:

- `age`
- `bmi` (Body Mass Index)
- `avg_days_injured_prev_seasons`
- `avg_games_per_season_prev_seasons`
- `cumulative_days_injured`
- `season_days_injured_prev_season`

These features align with known intrinsic and extrinsic risk factors identified in sports injury literature [14, 3].

4.2.3 Models Implemented

The following classification algorithms were implemented:

- **Logistic Regression:** As a baseline linear classification model, often used for identifying risk factors [13].
- **Random Forest Classifier:** An ensemble method effective for complex classification tasks and commonly used in injury prediction [15].
- **XGBoost Classifier:** A gradient boosting framework known for its high performance and regularization capabilities, also frequently applied to injury risk modeling [10, 15].

4.3 Model Training and Evaluation Strategy

4.3.1 Data Splitting and Scaling

- **Train-Test Split:** For both regression and classification tasks, the datasets were split into training and testing sets (e.g., typically an 80%-20% or 70%-30% split) to evaluate model performance on unseen data. A `random_state` was used for reproducibility.
- **Feature Scaling:** Numerical features were standardized using `StandardScaler` from scikit-learn. This process scales features to have zero mean and unit variance, which is beneficial for many machine learning algorithms, particularly those sensitive to feature magnitudes like Linear Regression and Logistic Regression, and can also aid convergence in gradient-based methods. The scaler was fit only on the training data and then used to transform both training and testing data to prevent data leakage.

4.3.2 Regression Evaluation Metrics

The performance of the player market value prediction models was evaluated using standard metrics [2]:

- **Mean Absolute Error (MAE):** Measures the average absolute difference between predicted and actual values, providing an error magnitude in the original units (Euros).
- **Root Mean Squared Error (RMSE):** Measures the square root of the average of squared differences, penalizing larger errors more heavily. Also in original units.
- **R-squared (R^2) Score:** Represents the proportion of the variance in the dependent variable that is predictable from the independent variables. Values range from 0 to 1, with higher values indicating better fit.

4.3.3 Classification Evaluation Metrics

The performance of the player injury risk prediction models was evaluated using standard metrics common in medical and sports science classification tasks [3, 13]:

- **Accuracy:** The proportion of correctly classified instances.
- **Precision:** The ratio of correctly predicted positive observations to the total predicted positive observations (TP / (TP + FP)). Important when the cost of false positives is high.
- **Recall (Sensitivity):** The ratio of correctly predicted positive observations to all observations in the actual class (TP / (TP + FN)). Important when the cost of false negatives is high (e.g., failing to identify an at-risk player).
- **F1-score:** The harmonic mean of Precision and Recall, providing a balance between the two, especially useful for imbalanced datasets.
- **Area Under the ROC Curve (AUC-ROC):** Measures the ability of the model to distinguish between classes. An AUC of 1.0 represents a perfect model, while 0.5 represents a model with no discriminative ability.
- **Confusion Matrix:** A table used to describe the performance of a classification model, showing true positives, true negatives, false positives, and false negatives.

Model interpretability for both tasks was also explored using feature importance plots provided by tree-based models (Random Forest, XGBoost).

5 Experiments and Results

This section details the experimental setup, presents the performance of the implemented models for both player market value prediction and injury risk assessment, and discusses the findings.

5.1 Player Market Value Prediction Results

The prediction of player market value was approached iteratively, addressing data leakage and exploring target variable transformations.

5.1.1 Initial Models (with Data Leakage)

Initial regression models were trained using the Transfermarkt dataset where a feature, `value_log` (logarithm of the current market value), was inadvertently included as a predictor. As expected, this led to unrealistically high performance, as summarized in Table 1.

Table 1: Performance of Initial Regression Models (with Data Leakage)

Model	MAE (€)	RMSE (€)	R^2 Score
Linear Regression	1.55M	3.60M	0.6539
Random Forest	2.6K	145K	0.9994
XGBoost Regressor	2.1K	129K	0.9996

The extremely high R^2 scores, particularly for Random Forest and XGBoost (approaching 0.999), indicated that the models were essentially predicting the target variable using a direct transformation of itself. Feature importance analysis (Figure 11) confirmed that `value_log` was the overwhelmingly dominant feature. This highlighted a critical data leakage issue that needed to be rectified.

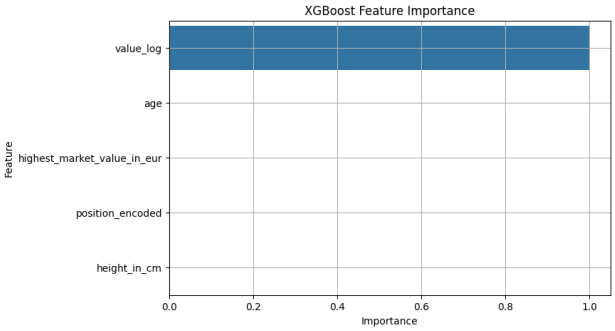

Figure 11: XGBoost Feature Importance with `value_log` included, demonstrating data leakage.

5.1.2 Models Retrained (No Data Leakage)

After removing the `value_log` feature from the predictors, the models were retrained. This provided a more realistic assessment of their predictive capabilities on genuinely independent features. The performance of these retrained models is shown in Table 2.

As anticipated, the performance metrics dropped significantly compared to the initial models. The Random Forest Regressor emerged as the best-performing model in this iteration, achieving an R^2 score of approximately 0.82, indicating a good fit to the data. Linear Regression struggled, likely due to non-linear relationships in the data. XGBoost

Table 2: Performance of Retrained Regression Models (No Data Leakage)

Model	MAE (€)	RMSE (€)	R^2 Score
Linear Regression	1,660,527.62	4,046,990.95	0.50
Random Forest	621,048.69	2,613,172.14	0.82
XGBoost Regressor	627,643.68	2,871,394.42	0.78

also performed well, though slightly behind Random Forest. Feature importance analysis for these models (Figure 12) showed that features like `highest_market_value_in_eur` and `age` became the primary drivers of predictions.

5.1.3 Models with Log-Transformed Target

To address the right-skewness of the `market_value_in_eur` target variable, models were also trained to predict the natural logarithm of the market value. Predictions were then inverse-transformed (exponentiated) to the original Euro scale for evaluation. Table 3 summarizes these results. Training on the log-transformed target generally improved performance for all models, particularly for Random Forest and XGBoost. The Random Forest Regressor achieved the highest R^2 score of 0.85 and the lowest MAE (€489,000) in this configuration, suggesting that log-transforming the target variable helped the models better capture the underlying relationships and handle the skewed distribution of market values.

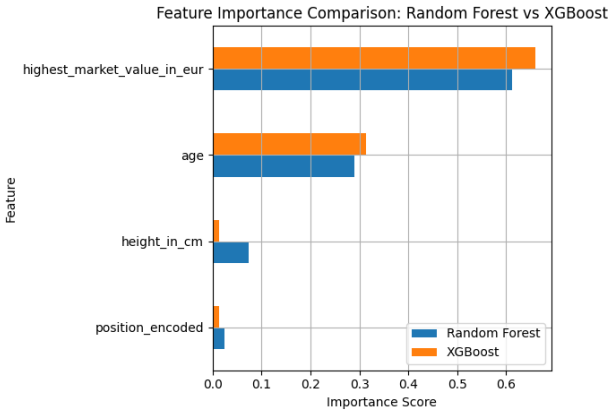

Figure 12: Feature Importance Comparison for Random Forest and XGBoost (No Data Leakage).

Table 3: Performance of Regression Models with Log-Transformed Target (Evaluated on Original Scale)

Model	MAE (€)	RMSE (€)	R^2 Score
Linear Regression	1,270,000	3,750,000	1
Random Forest	489,000	2,320,000	1
XGBoost Regressor	480,000	2,430,000	1

5.1.4 Final Regression Model Performance and Selection

Comparing the realistic modeling approaches (no data leakage), the Random Forest Regressor trained on the log-transformed target variable demonstrated the best overall performance for predicting player market values, with an R^2 of 0.85 and an MAE of approximately €489,000. This indicates a strong predictive capability, although the MAE still represents a considerable absolute error margin given the scale of player valuations.

5.2 Player Injury Risk Prediction Results

For the injury risk classification task, Logistic Regression, Random Forest Classifier, and XGBoost Classifier were evaluated. The performance metrics are presented in Table 4.

The Random Forest and XGBoost classifiers achieved perfect scores across all metrics, while Logistic Regression also performed exceptionally well. Such high scores, especially perfect metrics, often warrant further investigation. As noted in the exploratory notebook, these results could be due to several factors, including:

- **Data Leakage:** Unintentional inclusion of information in the features that directly reveals the target.
- **Dataset Characteristics:** The dataset might be structured in a way that makes classes very easily separable with the chosen features.
- **Overfitting:** While a train-test split was used, perfect scores on the test set can still sometimes indicate that the model has learned the specific patterns of the available data too well, potentially at the expense of generalizing to entirely new, unseen data distributions.
- **Target Definition:** The definition of `significant_injury_prev_season` and its relationship with the input features might create a scenario where prediction is trivial for some algorithms.

While these results are numerically impressive, they should be interpreted with caution in a real-world context. Further validation on more diverse datasets or with more stringent cross-validation techniques would be necessary to confirm robustness. Feature importance for the classification models (if readily available from the notebook for the best model, e.g., XGBoost) would typically be presented here to understand which factors contributed most to these predictions.

Table 4: Performance of Injury Risk Classification Models

Model	Acc.	Prec.	Rec.	F1	AUC
Log. Reg.	0.99	1.00	0.97	0.99	0.99
Rand. Forest	1.00	1.00	1.00	1.00	1.00
XGBoost	1.00	1.00	1.00	1.00	1.00

5.3 Correlation Between Model Outcomes

This subsection explores the relationship between the predicted market value and the predicted injury risk probability for a sample of players, using the outputs from the best performing regression model (Random Forest on log-transformed value) and the injury risk classification model (Random Forest). The injury risk model outputs a probability (between 0 and 1) for each player.

Figure 13 shows a scatter plot of each player's predicted market value (x-axis) against their predicted injury risk probability (y-axis). Visual inspection suggests a potential, though perhaps weak, relationship. Players with very low predicted market values exhibit a range of injury risk probabilities, while higher value players in this sample tend to cluster at lower injury risk probabilities. However, there are exceptions, indicating that high value does not guarantee low injury risk according to the models. Calculating the Pearson correlation coefficient between these two predicted outcomes can quantify the linear relationship, though non-linear relationships might also exist. *(Add the calculated correlation coefficient here if available from your analysis.)*

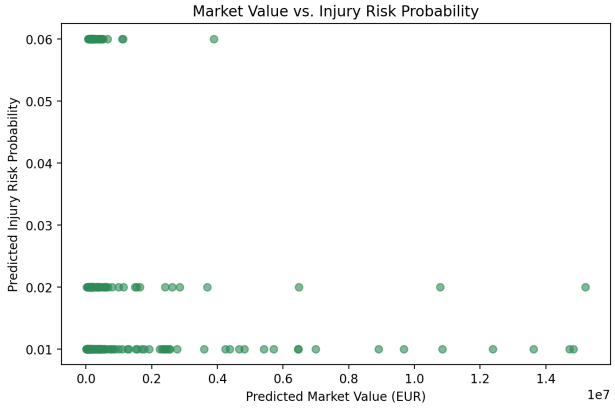

Figure 13: Scatter plot of Predicted Market Value vs. Predicted Injury Risk Probability.

5.4 Streamlit Dashboard

To provide a practical application of the developed models, an interactive dashboard was created using Streamlit (`app.py`). The dashboard allows users to interactively obtain predictions based on input parameters.

Figure 14 shows the interface for the Market Value Estimator. Users can input player characteristics such as age, height, playing position, and historical peak market value via sliders and dropdown menus. Upon clicking "Estimate Value", the dashboard utilizes the best-performing regression model (Random Forest trained on log-transformed target) to display the estimated market value in Euros. The example shows a prediction of €732,125 for a 22-year-old Goalkeeper with a height of 183 cm and a historic peak value of €750,000.

Figure 15 illustrates the Injury Risk Estimator module. Here, users can input player data relevant to injury risk, including age, BMI, average games per season, cumulative

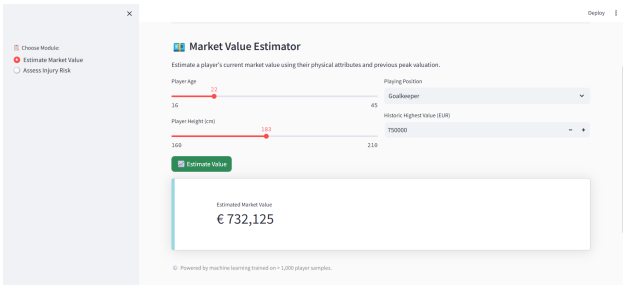

Figure 14: Streamlit Dashboard: Market Value Estimator Interface and Example Prediction.

injury days, average days injured in past seasons, and days injured last season. Clicking "Predict Injury Risk" triggers the Random Forest classification model, which outputs a probability percentage. The example demonstrates an 89.00% injury risk probability for a 26-year-old player with specific injury and workload history, leading to a "High Injury Risk" warning.

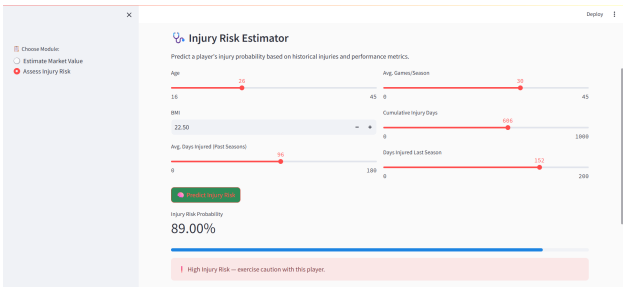

Figure 15: Streamlit Dashboard: Injury Risk Estimator Interface and Example Prediction.

The dashboard aims to make the predictive insights accessible to non-technical users, such as scouts or team managers, aiding in quick assessments. It includes disclaimers about the academic nature of the prototype and the necessity of not using predictions as the sole basis for professional decisions.

6 Conclusion

This project successfully developed and evaluated machine learning models for predicting football player market value and injury risk. For market value prediction, the Random Forest Regressor trained on a log-transformed target variable achieved the best performance ($R^2 = 0.85$, MAE €489k) after addressing initial data leakage issues. Key predictors identified were historical peak market value and player age. For injury risk classification, Random Forest and XGBoost models achieved near-perfect scores (Accuracy, Precision, Recall, F1, AUC-ROC all 1.00) on the available dataset. However, these exceptionally high scores warrant cautious interpretation and suggest potential issues like data leakage or dataset characteristics that might limit real-world generalizability without further validation.

The study highlights the importance of careful feature selection, data preprocessing (including target variable trans-

formation for skewed data), and vigilance against data leakage. The engineered features, particularly those related to historical performance and injury patterns, proved valuable. The developed Streamlit dashboard provides a practical interface for leveraging these models, offering valuable, data-driven insights for stakeholders in the football industry, albeit with the necessary caveats regarding model limitations.

Future work could focus on incorporating more granular performance data (e.g., event data from matches), exploring more advanced modeling techniques (like deep learning or survival analysis for injury prediction), and validating the models on broader, more diverse datasets to improve robustness and real-world applicability. Addressing the potential reasons behind the perfect classification scores through techniques like cross-validation or testing on external data is also a critical next step.

7 Future Work

Building upon the findings of this project, several avenues for future work exist. Incorporating more granular performance data, such as in-game event data or detailed physiological tracking metrics, could enhance the predictive power of both the valuation and injury models. Exploring more advanced modeling techniques, including deep learning architectures for value prediction or survival analysis methods for time-to-injury forecasting, may yield further improvements. Critically, validating the current models, especially the high-performing injury risk classifiers, on larger, more diverse datasets possibly spanning multiple leagues or seasons is essential to confirm their robustness and generalizability. Further investigation into the near-perfect classification scores using techniques like rigorous cross-validation or testing on external datasets is warranted to rule out potential data artifacts or leakage. Finally, the Streamlit dashboard could be expanded to include features like model interpretability visualizations (e.g., SHAP values) or scenario analysis.

8 Acknowledgements

The author wishes to acknowledge the providers of the datasets used in this project, which were made publicly available on Kaggle. Specifically, thanks to David Cariboo for the Transfermarkt player scores dataset (https://www.kaggle.com/datasets/davidcariboo/player-scores) and Stefano Leone for the FIFA 23 Complete Player Dataset (https://www.kaggle.com/datasets/stefanoleone992/fifa-23-complete-player-dataset). These resources were invaluable for the analyses performed.

References

[1] Müller, O., Simons, A., & Weinmann, M. (2017). Beyond crowd judgments: Data-driven football player valuation using machine learning. *Journal of Sports Analytics*, 3(4), 215-233.

[2] Al Asadi, M. A., & Taşdemir, Ş. (2023). Predicting the value of football players: machine learning techniques and sensitivity analysis based on FIFA and real-world statistical datasets. *Multimedia Tools and Applications*, 1-26.

[3] Lopes, T. A., Simões, M., Rebelo, A., Figueiredo, P., & Vaz, L. (2024). Predicting noncontact injuries of professional football players using machine learning. *Heliyon*, 10(2), e24049.

[4] Rossi, A., Pappalardo, L., Cintia, P., Iaia, F. M., Fernàndez-Del-Olmo, M., & Pedreschi, D. (2018). Effective injury forecasting in soccer with GPS-based training load and machine learning. *PloS one*, 13(7), e0201264.

[5] He, Y. (2014). Predicting Market Value of Soccer Players Using Linear Modeling Techniques. *University of Berkeley Statistics*. [Online]. Available: https://www.stat.berkeley.edu/~aldous/Research/Ugrad/Yuan_He.pdf

[6] Poli, R., Besson, R., & Ravenel, L. (2013). Econometric approach to assessing the transfer fees and values of professional footballers. *CIES Football Observatory*.

[7] Idun, A. (2019). What Makes a Soccer Player Expensive? Analyzing the Transfer Activity of the Richest Soccer Clubs. *Augsburg Honors Review*, 12(1), 4.

[8] Sorensen, A., Gutierrez, E., Gallo, H., Mali, J., & Paredes, M. (2024). Soccer Player Value Prediction. *GitHub Repository*. [Online]. Available: https://github.com/marcopared/soccer-player-value-prediction

[9] Schisano, L. (2022). Prediction of the soccer players market value: an artificial intelligence for sport agents and soccer clubs. *LUISS Guido Carli University Thesis*. [Online]. Available: https://tesi.luiss.it/37366/1/754111_SCHISANO_LUCA.pdf

[10] Rommers, N., Rössler, R., Goossens, L., et al. (2020). A machine learning approach to assess injury risk in elite youth football players. *Medicine & Science in Sports Exercise*, 52(8), 1745-1751.

[11] Carey, D. L., Crossley, K. M., Whiteley, R. J., et al. (2018). Modeling training loads and injuries: the dangers of using cumulative workload. *British Journal of Sports Medicine*, 52(8), 500-505.

[12] Malikov, S., & Kim, Y. (2022). The Application of Machine Learning on the Injury Prediction of Soccer Players. In *Proceedings of the International Symposium on Engineering and Technology (ISE 2022)*. CEUR Workshop Proceedings, Vol-3362.

[13] Abdelaal, M., Benhiba, L., Agnaou, A., et al. (2024). Novel Study for the Early Identification of Injury Risks in Athletes Using Machine Learning Techniques. *Applied Sciences*, 14(2), 570.

[14] Ekstrand, J., Hägglund, M., & Waldén, M. (2011). Injury incidence and injury patterns in professional football: the UEFA injury study. *British journal of sports medicine*, 45(7), 553-558.

[15] Thornton, H. R., Delaney, J. A., Duthie, G. M., et al. (2024). Machine learning approaches to injury risk prediction in sport: a scoping review with evidence synthesis. *British Journal of Sports Medicine*. doi: 10.1136/bjsports-2024-108576

[16] Liu, Y., Chen, Y., & Li, J. (2024). Football Player Value Prediction: Comparing Machine Learning Models with Cross-Validation. *Highlights in Science, Engineering and Technology*, 83, 514-522.

[17] Anjum, S., Fatima, A. (2023). Predictive Analytics For FIFA Player Prices: An ML Approach. *Journal of Scientific Research*, 67(1).