CAPSTONE FINAL PROJECT

**FOOTBALL PLAYERS MARKET PRICE PREDICTION**

**Course code:** BA723

**Course name:** Business Analytics Capstone

**Student’s name:** Marcos Edú Muñoz Quintasi

**Program:** Business Analytics and Insights – Canadian Context

**August 2023**

# ABSTRACT

This project aimed to predict football player market values by leveraging a comprehensive dataset encompassing key player attributes and performance metrics. Through rigorous analysis and modeling, our objective was to develop an accurate predictive framework to assist stakeholders in making informed decisions within the football industry.

The project's culmination led us to the selection of the Optimized Gradient Boosting Regressor as the best-performing model. With an impressively low Root Mean Squared Error (RMSE) of 0.65 and a high R-squared (R2) value of 0.78, the model demonstrated exceptional predictive capabilities. Its ability to capture intricate relationships between variables and player market values underscores its reliability and efficacy as a predictive tool.

Based on the insights gleaned from this endeavor, we offer the following recommendations to enhance decision-making within the football domain:

* Incorporate Youth Talent Analysis
* Strategize for Attacking Players
* Optimize Contract Negotiations

In conclusion, this project's successful application of the Optimized Gradient Boosting Regressor provides a valuable tool for predicting soccer player market values. By heeding our recommendations, stakeholders can leverage data-driven insights to strategize effectively, capitalize on market dynamics, and make well-informed decisions to drive success within the soccer industry.

**TABLE OF CONTENTS**

[ABSTRACT 2](#_Toc143006615)

[INTRODUCTION 6](#_Toc143006616)

[THE DATA 7](#_Toc143006617)

[Data Sources 7](#_Toc143006618)

[Data Workflow 8](#_Toc143006619)

[Data Gathering 8](#_Toc143006620)

[Data Cleaning 9](#_Toc143006621)

[Exploratory Data Analysis 11](#_Toc143006622)

[MODELLING 15](#_Toc143006623)

[Base Modelling 16](#_Toc143006624)

[Optimizing Models Using Hyperparameters 17](#_Toc143006625)

[Model Comparison 18](#_Toc143006626)

[Model Selection 18](#_Toc143006627)

[Most Important Features For Best Model 19](#_Toc143006628)

[CONCLUSION 21](#_Toc143006629)

[RECOMMENDATIONS 22](#_Toc143006630)

[APPENDICES 23](#_Toc143006631)

[REFERENCES 46](#_Toc143006632)

**TABLES**

[Table 1: Data Dictionary 23](#_Toc142984695)

[Table 2: Missing Values 30](#_Toc142984696)

[Table 3: Player Market Value Bin 32](#_Toc142984697)

[Table 4: Numerical features describe table 32](#_Toc142984698)

[Table 5: Numerical features describe table after Handling Outliers 33](#_Toc142984699)

**FIGURES**

[Figure 1: Histogram of Player Market Value 35](#_Toc142984700)

[Figure 2: Box Plot of Player Market Value 35](#_Toc142984701)

[Figure 3: Histogram of Log Transformed Player Market Value 35](#_Toc142984702)

[Figure 4: Bar Chart for Categorical Features 36](#_Toc142984703)

[Figure 5: Histogram for Numerical Features 37](#_Toc142984704)

[Figure 6: Box Plot for Numerical Features 40](#_Toc142984705)

[Figure 7: Heatmap for Numerical Subset 1 43](#_Toc142984706)

[Figure 8: Heatmap for Numerical Subset 2 43](#_Toc142984707)

[Figure 9: Heatmap for Numerical Subset 3 43](#_Toc142984708)

[Figure 10: Heatmap for Numerical Subset 4 43](#_Toc142984709)

[Figure 11: Heatmap for Numerical Subset 5 44](#_Toc142984710)

[Figure 12: Heatmap for Numerical Subset 6 44](#_Toc142984711)

[Figure 13: Heatmap for Numerical Subset 7 44](#_Toc142984712)

[Figure 14: Heatmap for Numerical Subset 8 45](#_Toc142984713)

[Figure 15: Heatmap for Numerical Subset 9 45](#_Toc142984714)

# INTRODUCTION

Football market value refers to the estimated worth of a player in the transfer market, indicating their demand and potential impact on a team. According to the most recent data, top-tier players’ market values are continuing to rise as a result of expanded media agreements, corporate sponsorships, and the sport's globalisation. Superstar player transfers have broken records, highlighting the game's economic importance and the continued global love with football.

The project aims to predict the market value of football male players during the 2022-2023 season in the top 5 leagues: Premier League (England), La Liga (Spain), Bundesliga (Germany), Serie A (Italy), Ligue 1 (France).

The project's inception stemmed from two driving factors. Firstly, a personal passion for football ignited the endeavor. Secondly, recognizing the vital role of data in the world of football to achieve sporting and financial goals. Using machine learning algorithms to boost decision-making will allow clubs to make data-driven player recruitment choices, enhance overall performance on the field, among other things.

Research questions that the project study seeks to answer:

* How does a player's market value vary across different positions and playing styles in football?
* Which specific player attributes and performance metrics have the most significant impact on determining market value in the football industry?
* How accurately can we predict a player's market value based on their attributes and season statistics?

# THE DATA

## Data Sources

Obtaining reliable and comprehensive data is a critical aspect of any data analytics research project. In this study, we explored two primary sources for football-related data: the Transfermarkt[[1]](#footnote-1) website and the FBref[[2]](#footnote-2) website. These platforms offer extensive datasets on player market values, performance statistics, and other essential information that can significantly enrich our analysis. Leveraging the vast resources available on these platforms will enable us to gain deeper insights into player valuations, player facts, and player performance.

Transfermarkt Website:

The Transfermarkt website stands as a prominent online platform for football-related data, focusing primarily on player market values, transfer news, and market trends. With a vast database covering numerous football leagues and players from around the globe, Transfermarkt provides an extensive collection of historical and real-time data. Their player market value estimations, derived from a combination of expert analysis and market trends, offer valuable insights into the dynamic nature of player valuations.

FBref Website:

FBref, a reputable sports data provider, specializes in presenting comprehensive and detailed statistics for football matches, players, and teams. The website's coverage spans various football leagues and competitions, ensuring access to diverse data points for in-depth analysis. Utilizing FBref's extensive data, we can conduct thorough assessments of player and team performances, identify key strengths and weaknesses, and develop models to predict player market values based on their on-field contributions and overall impact.

## Data Workflow

## Data Gathering

In this study, the data gathering involved a systematic approach to retrieve and compile relevant data from Transfermarkt and FBRef sources. The following steps were undertaken to construct the comprehensive dataset used for the project:

1. Retrieve Team's Transfermarkt Links:

To initiate the data collection process, Python libraries like BeautifulSoup and web scraping techniques were employed. The objective was to gather the Transfermarkt links of teams that participated in the 2022-2023 football season. This step served as the foundation for obtaining individual player information associated with each team.

1. Retrieve Player’s Name and Market Value:

Building on the previously obtained team links, the next step focused on acquiring player-specific data. The Python script utilized the team links to extract player names and their corresponding market values for the 2022-2023 season from the Transfermarkt website. These market values represent crucial data points for player valuation and served as the output variable for the project.

1. Get Players Main Facts:

To complement the player dataset with additional relevant information, the worldfootballR and dplyr libraries were utilized using RStudio. This allowed us to iterate through the Transfermarkt links of each team and gather players' main facts, such as height, age, and date of birth. These demographic attributes contribute significantly to understanding player profiles.

1. Get Player Season Statistics:

Using again the worldfootballR and dplyr libraries in RStudio, detailed player statistics for the 2022-2023 season were collected. A function inside the worldfootballR package was employed to obtain various performance metrics, including standard statistics, shooting data, passing insights, defensive contributions, and playing time, among others. These statistics provide a comprehensive overview of player performance during the season.

1. Handling Players in Multiple Teams:

Some players may have participated in more than one team during the season, necessitating careful handling to avoid duplication and data integrity issues. To address this, player statistics were aggregated using sum or average methods, depending on the nature of the variable. This ensured that each player's performance was accurately represented in the dataset.

1. Generate Raw Dataset:

By performing a left join on the Transfermarkt player links and FBref player links, the raw dataset was created, integrating information from both sources. Additionally, duplicate entries were efficiently handled during this process, ensuring that the raw dataset only contained unique players. The resulting raw final dataset was saved as a CSV file, ready for further data cleaning and model building. The raw dataset contains 3555 rows and 113 columns.

## Data Cleaning

The data cleaning process played a crucial role in ensuring the integrity and accuracy of the dataset. The following steps were undertaken to clean the raw data and prepare it for exploratory data analysis:

*Load DataFrame:*

The first step involved loading the raw final dataset, which was generated after merging the Transfermarkt and FBref data sources, into a pandas DataFrame. This allowed for easy manipulation and transformation of the data.

*Remove Rows with Incomplete Information:*

Rows containing incomplete information posed a challenge to the analysis. To address this, all players without a market value, players with missing season statistics, and players with no matches played during the 2022-2023 season were removed from the DataFrame. This ensured that only players with sufficient data for analysis remained in the dataset.

*Remove Unnecessary Columns:*

Certain columns in the DataFrame were identified as redundant or unnecessary for the analysis. Specifically, the "current club" column was removed as it duplicated information already present in other columns, and the "player name" column was deemed unnecessary due to the existence of an identification (ID) column.

*Handle Missing Values:*

A DataFrame was created to visualize the percentage of missing values per column (Table 2: Missing Values), aiding in identifying data gaps. Goalkeeper players and goalkeeper statistics were removed, as the focus was on outfield players. The "OUTFITTER" column was transformed to a boolean column, representing True if the value was not null, indicating whether the player had an outfitter. The "DIST\_STANDARD" column was imputed with the median value to address missing values in player distance statistics. Similarly, the "PLAYER\_AGENT" column was transformed into a boolean column, indicating whether the player had an agent. Missing values in the "HEIGHT" column were imputed with the median height value, while the "FOOT" column was imputed with the value "right" to handle missing foot preference information.

*Convert String Values to Numeric:*

To facilitate analysis, a function was created to convert string values in the "PLAYER\_VALUE" column, representing player market values, into numeric format. This transformation was required to build the regression model.

The culmination of these data cleaning efforts resulted in a cleaned DataFrame, containing 2384 rows and 94 columns, with consistent and well-structured data ready for subsequent data analytics tasks. The cleaned dataset was saved as a CSV file named "cleaned\_dataset.csv," paving the way for further exploration and analysis.

## Exploratory Data Analysis

The Exploratory Data Analysis phase played a pivotal role in gaining insights into the dataset and preparing it for subsequent modeling. The following steps were undertaken to explore, visualize, and preprocess the data:

*Import Packages and Load the Data:*

The EDA journey commenced by importing essential libraries and loading the cleaned dataset into the analysis environment. This step set the stage for comprehensive exploration and manipulation.

*Change Data Types:*

Data types were strategically adjusted for efficient analysis. Object columns and 'year\_birth' were converted into categorical variables, optimizing memory usage and enhancing interpretability.

*Visualize Output Variable:*

The analysis delved into the player market values, the output variable of interest. The player market values are listed in Euros (€) and the range goes from €500,000 to €180,000,000. Bins with a length of €25,000,000 were created to categorize market values and have a quick look about the distribution. Almost 90% of players (see Table 3) has a 'PLAYER\_VALUE' less than €25,000,000 and just 2 players (0.008% of players) have a ‘PLAYER\_VALUE’ above €175,000,000.

Then, visualizations were plotted illustrating the distribution through a histogram (see Figure 1) and a box plot (see Figure 2). The histogram and box plot reveal a notable right-skewness in the player market value distribution, indicating a concentration of values towards the lower end of the scale. Additionally, the presence of numerous outliers beyond the upper whisker in the box plot suggests significant variations in player market values, potentially influenced by exceptional factors or high-performing individuals.

The log transformation was applied to the output variable, player market value, to mitigate the right-skewness observed in its distribution. This transformation effectively compresses higher values while enhancing the visibility of lower values, aligning the variable more closely with normality (see Figure 3). By reducing the impact of extreme outliers, the log transformation contributes to a more balanced and representative feature for subsequent modeling analysis.

*Visualize Features:*

The exploration extended to the dataset's features. Notably, 'MACRO\_POSITION' was derived from the 'POSITION' variable, categorizing players into attack, midfield, or defense roles. Functions were created to generate bar plots for categorical features, histograms for numerical features, and box plots for feature distributions. Furthermore, a strategic subset of numerical features was selected based on correlation, and correlated variables with lower variance were dropped. In total, there was 45 correlated variables that were dropped.

*Create Dummies for Categorical Features:*

Label encoding and one-hot encoding are crucial techniques in data preprocessing, particularly when predicting a continuous variable. Label encoding involves converting categorical variables into numerical values, assigning a unique integer to each category. This simplifies the data representation, allowing algorithms to process categorical information.

One-hot encoding, on the other hand, transforms categorical variables into binary columns, where each column represents a distinct category, indicating its presence or absence. These encoding methods are essential as continuous variable prediction models, such as regression, rely on numerical input. By encoding categorical features, we ensure compatibility with these models, preventing biases or misinterpretations. Moreover, one-hot encoding prevents false ordinal relationships among categories, crucial when categorical variables do not possess inherent order.

Categorical features underwent transformation for enhanced model compatibility. 'CURRENT\_INTERNATIONAL,' 'CLUB\_NAME,' and 'YEAR\_BIRTH' were label-encoded, while other categorical features underwent one-hot encoding, expanding the dataset's feature space.

*Handling Outliers*

A reliable strategy for identifying and reducing the impact of extreme data points within a dataset is to handle outliers using the Z-score method (see Table 4: Numerical features describe table). The Z-score quantifies how far a data point is from the average by measuring the deviation of each data point from the mean in terms of standard deviations. Each data point's Z-score is calculated, and outliers are those that have Z-scores higher than a predetermined cutoff (set at 3). A description table of numerical features was constructed after handling outliers to show the changes. (see Table 5: Numerical features describe table after Handling Outliers).

*Scale Numerical Features:*

MinMaxScaler is a vital data preprocessing technique that plays a pivotal role in preparing data for analysis and modeling. It ensures that all numerical features are transformed and scaled to a specific range, typically between 0 and 1. This standardization process is essential because it brings all variables to a common scale, mitigating issues arising from differences in the magnitude of feature values.

By doing so, MinMaxScaler prevents certain features from dominating others solely due to their larger values, thus helping models to converge faster and perform more effectively.

Additionally, MinMaxScaler is particularly useful when using algorithms that are sensitive to feature scaling, such as gradient-based optimization methods, distance-based algorithms, and neural networks. By applying MinMaxScaler as part of the preprocessing pipeline, we ensure that our data is appropriately scaled, enhancing the model's ability to capture underlying patterns, improve convergence, and produce more accurate predictions.

*Save Filtered and Scaled Features:*

The EDA phase concluded with the creation of a CSV file containing the filtered and scaled features, meticulously prepared for subsequent modeling endeavors. This step ensured that the data was organized, preprocessed, and ready to be fed into various machine learning algorithms.

By meticulously executing these steps, the Exploratory Data Analysis not only offered valuable insights into the dataset's characteristics but also paved the way for robust and effective modeling, enabling the extraction of meaningful patterns and relationships for predictive analysis.

# MODELLING

Regression, a fundamental statistical technique, involves establishing a mathematical relationship between input variables and the target output, enabling us to estimate values.

Leveraging regression algorithms, such as linear regression, ridge regression, or even advanced machine learning techniques like random forests or gradient boosting, empowers us to capture intricate relationships within the data. Regular model evaluation and refinement complete the iterative process, ensuring that the regression model attains a robust and reliable predictive capacity for Market Player Value.

An evaluate function was created for assessing and presenting the performance metrics of a predictive model. The function, named "evaluate\_model," takes two arguments: the true target values (y\_true) and the predicted values (y\_pred). Upon invocation, the function calculates the Root Mean Squared Error (RMSE) and the coefficient of determination (R-squared or R^2) using these input arrays. RMSE provides a measure of the average magnitude of the prediction errors, conveying how closely the predicted values align with the actual values. Meanwhile, R^2 offers insights into the proportion of variance in the target variable that the model explains.

One of the primary methods used in machine learning to evaluate model performance is the train-test split strategy. The training set and testing set are two subsets that are created by splitting the dataset in half. The model is trained using the training set, which enables it to discover trends and connections in the data. The testing set is used to assess the model's generalizability and predict data that has not yet been encountered by the model during training.

Using test\_size = 0.4 in this situation denotes a 40% allocation of the dataset to the testing set and a 60% allocation to the training set. By providing a specified seed for the random number generator, the argument random\_state = 1 ensures reproducibility. This makes sure that the split is the same each time the code is run, making it easier to consistently compare the performances of various models and adjust their parameters.

## Base Modelling

*Linear Regression Model*

A statistical method for modeling the relationship between a dependent variable and one or more independent variables using a straight-line equation, aiming to predict outcomes.

Linear Regression (Train Set):

RMSE: 0.83, R^2: 0.66

Linear Regression (Test Set):

RMSE: 0.83, R^2: 0.64

*Decision Tree Model*

A tree-like model in machine learning that recursively partitions data based on feature values, aiding classification, and regression tasks by hierarchically making decisions.

Decision Trees (Train Set):

RMSE: 0.00, R^2: 1.00

Decision Trees (Test Set):

RMSE: 1.05, R^2: 0.42

*Lasso Regression*

A regularization technique that enhances linear regression by adding a penalty term to the absolute values of coefficients, promoting feature selection and mitigating overfitting in high-dimensional datasets.

Lasso Regression (Train Set):

RMSE: 1.42, R^2: 0.00

Lasso Regression (Test Set):

RMSE: 1.37, R^2: 0.00

*Decision Tree Regressor Model*

A predictive model that employs a tree-like structure to partition data, predicting continuous numerical outcomes by recursively evaluating feature conditions within distinct branches of the tree.

Decision Tree Regression (Train Set):

RMSE: 0.00, R^2: 1.00

Decision Tree Regression (Test Set):

RMSE: 1.05, R^2: 0.42

K-fold cross-validation is a reliable method for assessing a machine learning model's performance while maximising data utilisation. It entails dividing the dataset into "k" folds of equal size. The model is subsequently trained and verified "k" times, with the first fold serving as the validation set and the subsequent folds acting as the training set.

The dataset is split into five subgroups when n\_splits = 5. One of these subsets is used as the validation set for each iteration, while the remaining four are combined to create the training set. This method guarantees that every piece of data is utilised exactly once for validation and four times for training. By providing a specific seed for the random number generator, using random\_state = 1 ensures reproducibility.

*Linear Regression Model*

Linear Regression (k-fold CV):

RMSE: 0.84, R²: 0.65

*Decision Tree Model*

Decision Trees (k-fold CV):

RMSE: 1.04, R²: 0.78

*Lasso Regression*

Lasso Regression (k-fold CV):

RMSE: 1.40, R²: 0.00

*Decision Tree Regressor Model*

Decision Tree Regression (k-fold CV):

RMSE: 1.04, R²: 0.78

## Optimizing Models Using Hyperparameters

Model optimization through hyperparameter tuning is a crucial process in machine learning aimed at enhancing a model's performance by fine-tuning its hyperparameters. Hyperparameters are parameters set prior to training that influence the learning process and affect the model's capacity to generalize patterns from data. Hyperparameter tuning involves systematically searching and selecting the optimal combination of hyperparameters to achieve the best model performance.

Hyperparameter tuning prevents overfitting by controlling model complexity and ensures the model is well-suited to the specific dataset. It is an iterative and computationally intensive process, often involving multiple rounds of experimentation. Tools like cross-validation help validate the model's performance during the tuning process. Hyperparameter tuning ultimately leads to improved model accuracy, robustness, and better predictive capabilities.

*Optimized Decision Tree Regressor*

Best Hyperparameters: {'max\_depth': 5, 'min\_samples\_leaf': 2, 'min\_samples\_split': 5}

Best RMSE: 0.91

Best R-squared: 0.56

Train RMSE: 0.79, Train R²: 0.69

*Optimized Linear Regression using Gradient Boosting Regressor*

Best Hyperparameters: {'learning\_rate': 0.2, 'max\_depth': 3, 'n\_estimators': 200}

Best RMSE: 0.6527793204881664

Best R-squared: 0.9304604540636039

## Model Comparison

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Train** | | **Test** | |
| **Model** | **RMSE** | **R2** | **RMSE** | **R2** |
| Linear Regression | 0.83 | 0.66 | 0.83 | 0.64 |
| Decision Tree | 0 | 1 | 1.05 | 0.42 |
| Lasso Regression | 1.42 | 0 | 1.37 | 0 |
| Decision Tree Regression | 0 | 1 | 1.05 | 0.42 |
| Optimized Decision Tree | 0.79 | 0.69 | 0.91 | 0.56 |
| Optimized Gradient Boosting Regressor | 0.37 | 0.93 | 0.65 | 0.78 |

## Model Selection

After an extensive process of model evaluation and optimization, the Optimized Gradient Boosting Regressor has emerged as the most favorable choice. With a Root Mean Squared Error (RMSE) of 0.65 and a commendable coefficient of determination (R-squared or R2) of 0.78, this model has demonstrated its superior predictive capacity.

The achieved RMSE signifies the minimal average magnitude of prediction errors, reflecting the model's accuracy in estimating target values. Moreover, the high R2 value of 0.78 underscores the model's ability to explain variance in the target variable, indicating a strong alignment between predicted and actual outcomes.

## Most Important Features For Best Model

In the realm of Gradient Boosting Regressors, the concept of feature importance assumes a pivotal role in unraveling the intrinsic relationships within complex datasets. Feature importance provides a holistic view of the contribution each input variable makes towards the model's predictive prowess. By leveraging techniques such as permutation importance or analyzing the impurity reduction within decision trees, the Gradient Boosting Regressor enables us to rank and prioritize features based on their relative significance.

A graph with blue and black text

Description automatically generated

Here are some of the most important features:

1. \*\*Total Carrying Distance\*\*: The total distance a player covers while controlling the ball with their feet is a dynamic measure of their dribbling skills, agility, and ability to penetrate opposing defenses. Players who exhibit higher total carrying distances often possess exceptional ball control and the capacity to create goal-scoring opportunities through calculated movements on the field.

2. \*\*Age\*\*: Age remains a cornerstone feature as it profoundly impacts a player's physical attributes, tactical acumen, and overall experience. Younger players might possess raw talent and potential for growth, while older players may bring seasoned skills and leadership qualities to the team.

3. \*\*Attacking Penalty Touches\*\*: The number of touches a player records in the attacking penalty area reflects their involvement in offensive plays near the goal. Players with a high count of attacking penalty touches are likely to be key contributors to goal-scoring opportunities, indicating their active participation in offensive strategies.

4. \*\*Goal Difference\*\*: The goal difference achieved by the team while the player is on the pitch signifies their impact on team performance. Positive goal differences highlight a player's contribution to offensive play and defensive stability, showcasing their ability to both score and defend effectively.

5. \*\*Minutes Played\*\*: The total minutes a player spends on the field serves as a proxy for their consistency and durability. Players with significant playing time often possess the trust of coaches, implying their importance within the team's strategies and plans.

6. \*\*Club Name\*\*: The club a player belongs to can significantly influence their playing style, level of competition, and overall exposure. Players from renowned clubs may benefit from extensive resources, top-tier coaching, and higher visibility, which impact their performance and market value.

In conclusion, Total Carrying Distance, Age, Attacking Penalty Touches, Goal Difference, Minutes Played, and Club Name stand out as pivotal indicators of a player's prowess and potential impact on the prediction of the market value of the player.

# CONCLUSION

Intriguing insights were gleaned through the visual representation of market value against player positions. Our analysis revealed a notable trend: attacking players, characterized by their positions in the forward and midfield areas, tend to command higher market values compared to defensive players. This observation sheds light on the significance of offensive prowess and goal-scoring potential in influencing a player's perceived value. This finding resonates with the increasing emphasis on dynamic attacking play in modern football and highlights the market's appreciation for players who contribute significantly to a team's goal-scoring capabilities.

Employing the sophisticated Gradient Boosting Regressor model, we obtained impressive results in predicting player market values. Achieving a low Root Mean Squared Error (RMSE) of 0.65 and a commendable R-squared (R2) value of 0.78, the model demonstrated its proficiency in capturing the intricate relationships between the selected features and market values. This high predictive accuracy attests to the robustness of the model and its ability to offer informed estimates of player valuations, thus empowering stakeholders with a reliable tool for decision-making in the competitive football landscape.

The exploration and analysis of our dataset have revealed several standout variables that play a pivotal role in predicting a player's market value. Notably, Total Carrying Distance emerged as a key determinant, reflecting a player's ability to maneuver the ball skillfully and create scoring opportunities. Alongside this, Age showcased its significance, underscoring the influence of experience, physical attributes, and potential growth on a player's valuation. Moreover, Attacking Penalty Touches exhibited a strong correlation with market value, indicative of players who actively contribute to offensive plays near the goal. These variables, combined with Goal Difference and Minutes Played, offer a comprehensive framework for predicting player market value, providing valuable insights into the dynamic factors shaping football player valuations.

# RECOMMENDATIONS

Here are some recommendations to enhance the effectiveness and scope of the project:

*Data Collection and Preprocessing*

* Gather comprehensive data for other leagues to test the model in different players to avoid overfitting.
* Gather comprehensive data for promising players aged 14-18, including performance statistics, attributes, and potential indicators of market value.
* Acquire data on media coverage and social media metrics (followers, engagement) to capture the player's public visibility.
* Collate information on contract types, transfer history, and injury records for a holistic understanding of player dynamics.

*Continuous Monitoring and Updates:*

* Regularly update the model with fresh data to ensure its relevance and accuracy over time.
* Stay attuned to emerging trends and shifts in the football ecosystem that might influence market values.

By incorporating these recommendations, the project can offer insightful and actionable predictions for the market values of promising young players, contributing to better decision-making within the football industry.

# APPENDICES

**Tables**

Table : Data Dictionary

|  |  |  |  |
| --- | --- | --- | --- |
| **VARIABLE** | **NAME** | **EXPLANATION** | **AGGREGATION** |
| **PLAYER'S FACTS** | | | |
| CLUB\_NAME | Name of the club |  |  |
| PLAYER\_NAME | Name of the player |  |  |
| PLAYER\_VALUE | Current market value of the player | As of 2023-07-13 |  |
| PLAYER\_HREF | Player Transfermkt Link |  |  |
| CLUB\_HREF | Club Transfermkt Link |  |  |
| LEAGUE\_NAME | Player league name |  |  |
| LEAGUE\_COUNTRY | Player league country |  |  |
| UrlFBref | Player FBref Link |  |  |
| DATE\_OF\_BIRTH | Date of birth |  |  |
| AGE | Age |  |  |
| HEIGHT | Height |  |  |
| CURRENT\_INTERNATIONAL | Citizenship |  |  |
| CURRENT\_CLUB | Current club | Player's current club |  |
| JOINED | Joined | when does the player joined the current club |  |
| POSITION | Position |  |  |
| FOOT | Foot | Preferred foot (left, right, both) |  |
| PLAYER\_AGENT | Player agent | Player's representative |  |
| OUTFITTER | Outfitter | Company that wears the player |  |
| MAX\_PLAYER\_VALUATION | Max market value of the player |  |  |
| MAX\_PLAYER\_VALUATION\_DATE | Max market value valuation date |  |  |
| **STANDARD STATS** | | | |
| MP\_PLAYING | Matches played | Matches Played by the player or squad | SUM |
| STARTS\_PLAYING | Matches started | Game or games started by player | SUM |
| MIN\_PLAYING | Minutes played | Minutes played | SUM |
| GLS | Goals | Goals scored or allowed | SUM |
| AST | Assists | Assists | SUM |
| G+A | Goals + Assists | Goals and Assists | SUM |
| G\_MINUS\_PK | Non-Penalty Goals | Non-Penalty Goals | SUM |
| PK | Penalty Kicks Made | Penalty Kicks Made | SUM |
| PKATT | Penalty Kicks Attempted | Penalty Kicks Attempted | SUM |
| CRDY | Yellow Cards | Yellow Cards | SUM |
| CRDR | Red Cards | Red Cards | SUM |
| PRGC\_PROGRESSION | Progressive Carries | Carries that move the ball towards the opponent's goal line at least 10 yards from its furthest point in the last six passes, or any carry into the penalty area. Excludes carries which end in the defending 50% of the pitch | SUM |
| PRGP\_PROGRESSION | Progressive Passes | Completed passes that move the ball towards the opponent's goal line at least 10 yards from its furthest point in the last six passes, or any completed pass into the penalty area. Excludes passes from the defending 40% of the pitch | SUM |
| PRGR\_PROGRESSION | Progressive Passes Received | Completed passes that move the ball towards the opponent's goal line at least 10 yards from its furthest point in the last six passes, or any completed pass into the penalty area. Excludes passes from the defending 40% of the pitch | SUM |
| **SHOOTING STATS** | | | |
| SH\_STANDARD | Shots Total | Shots Total. Does not include penalty kicks | SUM |
| SOT\_STANDARD | Shots on target | Shots on target | SUM |
| DIST\_STANDARD | Average distance shot | Average distance, in yards, from goal of all shots taken | AVG |
| FK\_STANDARD | Shots from free kicks | Shots from free kicks | SUM |
| **POSSESSION STATS** | | | |
| TOUCHES\_TOUCHES | Touches | Number of times a player touched the ball. Note: Receiving a pass, then dribbling, then sending a pass counts as one touch | SUM |
| DEF PEN\_TOUCHES | Defensive penalty touches | Touches in defensive penalty area | SUM |
| DEF 3RD\_TOUCHES | Defensive third touches | Touches in defensive 1/3 | SUM |
| MID 3RD\_TOUCHES | Mid third touches | Touches in middle 1/3 | SUM |
| ATT 3RD\_TOUCHES | Attacking third touches | Touches in attacking 1/3 | SUM |
| ATT PEN\_TOUCHES | Attacking penalty touches | Touches in attacking penalty area | SUM |
| LIVE\_TOUCHES | Live touches | Live-ball touches. Does not include corner kicks, free kicks, throw-ins, kick-offs, goal kicks or penalty kicks | SUM |
| ATT\_TAKE | Take-Ons Attempted | Number of attempts to take on defenders while dribbling | SUM |
| SUCC\_TAKE | Successful Take-Ons | Number of defenders taken on successfully, by dribbling past them | SUM |
| TKLD\_TAKE | Times Tackled During Take-On | Number of times tackled by a defender during a take-on attempt | SUM |
| CARRIES\_CARRIES | Carries | Number of times the player controlled the ball with their feet | SUM |
| TOTDIST\_CARRIES | Total Carrying Distance | Total distance, in yards, a player moved the ball while controlling it with their feet, in any direction | SUM |
| PRGDIST\_CARRIES | Progressive Carrying Distance | Total distance, in yards, a player moved the ball while controlling it with their feet towards the opponent's goal | SUM |
| PRGC\_CARRIES | Progressive Carries | Carries that move the ball towards the opponent's goal line at least 10 yards from its furthest point in the last six passes, or any carry into the penalty area. | SUM |
| FINAL\_THIRD\_CARRIES | Carries into Final Third | Carries that enter the 1/3 of the pitch closest to the goal | SUM |
| CPA\_CARRIES | Carries into Penalty Area | Carries into the 18-yard box | SUM |
| MIS\_CARRIES | Miscontrols | Number of times a player failed when attempting to gain control of a ball | SUM |
| DIS\_CARRIES | Dispossessed | Number of times a player loses control of the ball after being tackled by an opposing player. Does not include attempted take-ons | SUM |
| REC\_RECEIVING | Passes Received | Number of times a player successfully received a pass | SUM |
| PRGR\_RECEIVING | Progressive Passes Rec | Completed passes that move the ball towards the opponent's goal line at least 10 yards from its furthest point in the last six passes, or any completed pass into the penalty area. | SUM |
| **PLAYING TIME STATS** | | | |
| COMPL\_STARTS | Complete matches played | Complete matches played | SUM |
| SUBS\_SUBS | Games Played as Sub | Games as sub. Game or games player did not start, so as a substitute | SUM |
| UNSUB\_SUBS | Games as unused sub | Games as an unused substitute | SUM |
| PPM\_TEAM.SUCCESS | Points per Match | Average number of points earned by the team from matches in which the player appeared | AVG |
| ONG\_TEAM.SUCCESS | Goals Scored (on pitch) | Goals scored by team while on pitch | SUM |
| ONGA\_TEAM.SUCCESS | Goals Allowed (on pitch) | Goals allowed by team while on pitch | SUM |
| PLUS\_PER\_\_MINUS\_\_TEAM.SUCCESS | Goal Difference | Goals scored minus goals allowed by the team while the player was on the pitch. | SUM |
| **PASSING STATS** | | | |
| CMP\_TOTAL | Passes Completed | Passes Completed | SUM |
| ATT\_TOTAL | Passes Attempted | Passes Attempted | SUM |
| TOTDIST\_TOTAL | Total Passing Distance | Total distance, in yards, that completed passes have traveled in any direction | SUM |
| PRGDIST\_TOTAL | Progressive Passing Distance | Total distance, in yards, that completed passes have traveled towards the opponent's goal. Note: Passes away from opponent's goal are counted as zero progressive yards. | SUM |
| KP | Key Passes | Passes that directly lead to a shot (assisted shots) | SUM |
| FINAL\_THIRD | Passes into Final Third | Completed passes that enter the 1/3 of the pitch closest to the goal | SUM |
| PPA | Passes into Penalty Area | Completed passes into the 18-yard box | SUM |
| CRSPA | Crosses into Penalty Area | Completed crosses into the 18-yard box | SUM |
| **MISC STATS** | | | |
| 2CRDY | Second yellow card |  | SUM |
| FLS | Fouls Committed |  | SUM |
| FLD | Fouls Drawn |  | SUM |
| OFF | Offsides |  | SUM |
| CRS | Crosses |  | SUM |
| TKLW | Tackles Won | Tackles in which the tackler's team won possession of the ball | SUM |
| PKWON | Penalty Kicks Won |  | SUM |
| PKCON | Penalty Kicks Conceded |  | SUM |
| OG | Own Goals |  | SUM |
| RECOV | Number of loose balls recovered |  | SUM |
| WON\_AERIAL | Aerials won |  | SUM |
| LOST\_AERIAL | Aerials lost |  | SUM |
| **KEEPER STATS** | | | |
| GA | Goals Against |  | SUM |
| SOTA | Shots on Target Against |  | SUM |
| SAVES | Number of saves |  | SUM |
| W | Wins |  | SUM |
| D | Draws |  | SUM |
| L | Losses |  | SUM |
| CS | Clean Sheets | Full matches by goalkeeper where no goals are allowed. | SUM |
| PKATT\_PENALTY | Penalty Kicks Attempted |  | SUM |
| PKA\_PENALTY | Penalty Kicks Allowed |  | SUM |
| PKSV\_PENALTY | Penalty Kicks Saved |  | SUM |
| PKM\_PENALTY | Penalty Kicks Missed |  | SUM |
| **GCA STATS** | | | |
| SCA\_SCA | Shot-creating Actions | The two offensive actions directly leading to a shot, such as passes, take-ons and drawing fouls. Note: A single player can receive credit for multiple actions and the shot-taker can also receive credit. | SUM |
| GCA\_GCA | Goal-creating Actions | The two offensive actions directly leading to a goal, such as passes, take-ons and drawing fouls. Note: A single player can receive credit for multiple actions and the shot-taker can also receive credit. | SUM |
| **DEFENSE STATS** | | | |
| TKL\_TACKLES | Tackles | Number of players tackled | SUM |
| TKLW\_TACKLES | Tackles won | Tackles in which the tackler's team won possession of the ball | SUM |
| DEF 3RD\_TACKLES | Tackles defensive 1/3 | Tackles in defensive 1/3 | SUM |
| MID 3RD\_TACKLES | Tackles mid 1/3 | Tackles in middle 1/3 | SUM |
| ATT 3RD\_TACKLES | Tackles attack 1/3 | Tackles in attacking 1/3 | SUM |
| TKL\_CHALLENGES | Dribblers Tackled | Number of dribblers tackled | SUM |
| ATT\_CHALLENGES | Dribblers Challenged | Number of unsuccessful challenges plus number of dribblers tackled | SUM |
| LOST\_CHALLENGES | Challenges Lost | Number of unsucessful attempts to challenge a dribbling player | SUM |
| BLOCKS\_BLOCKS | Blocks | Number of times blocking the ball by standing in its path | SUM |
| SH\_BLOCKS | Shots blocked | Number of times blocking a shot by standing in its path | SUM |
| PASS\_BLOCKS | Passes blocked | Number of times blocking a pass by standing in its path | SUM |
| INT | Interceptions | Interceptions | SUM |
| TKL+INT | Number of players tackled plus number of interceptions | Number of players tackled plus number of interceptions | SUM |
| CLR | Clearances | Clearances | SUM |
| ERR | Errors | Mistakes leading to an opponent's shot | SUM |

Table : Missing Values

|  |  |  |
| --- | --- | --- |
| **Variable** | **Missing Values** | **Missing Percentage** |
| CS | 2384 | 93% |
| D | 2384 | 93% |
| PKA\_PENALTY | 2384 | 93% |
| PKSV\_PENALTY | 2384 | 93% |
| L | 2384 | 93% |
| PKATT\_PENALTY | 2384 | 93% |
| PKM\_PENALTY | 2384 | 93% |
| W | 2384 | 93% |
| SAVES | 2384 | 93% |
| SOTA | 2384 | 93% |
| GA | 2384 | 93% |
| OUTFITTER | 1580 | 61% |
| DIST\_STANDARD | 424 | 16% |
| PLAYER\_AGENT | 248 | 10% |
| MAX\_PLAYER\_VALUATION | 210 | 8% |
| FOOT | 40 | 2% |
| HEIGHT | 27 | 1% |
| JOINED | 8 | 0% |
| DATE\_OF\_BIRTH | 8 | 0% |
| PKCON | 0 | 0% |
| TKLW | 0 | 0% |
| PKWON | 0 | 0% |
| FLS | 0 | 0% |
| CRS | 0 | 0% |
| OFF | 0 | 0% |
| FLD | 0 | 0% |
| RECOV | 0 | 0% |
| 2CRDY | 0 | 0% |
| CRSPA | 0 | 0% |
| PPA | 0 | 0% |
| FINAL\_THIRD | 0 | 0% |
| KP | 0 | 0% |
| PRGDIST\_TOTAL | 0 | 0% |
| TOTDIST\_TOTAL | 0 | 0% |
| ATT\_TOTAL | 0 | 0% |
| CMP\_TOTAL | 0 | 0% |
| PLUS\_PER\_\_MINUS\_\_TEAM.SUCCESS | 0 | 0% |
| OG | 0 | 0% |
| CLUB\_NAME | 0 | 0% |
| WON\_AERIAL | 0 | 0% |
| LOST\_AERIAL | 0 | 0% |
| CLR | 0 | 0% |
| TKL+INT | 0 | 0% |
| INT | 0 | 0% |
| PASS\_BLOCKS | 0 | 0% |
| SH\_BLOCKS | 0 | 0% |
| BLOCKS\_BLOCKS | 0 | 0% |
| LOST\_CHALLENGES | 0 | 0% |
| ATT\_CHALLENGES | 0 | 0% |
| TKL\_CHALLENGES | 0 | 0% |
| ATT 3RD\_TACKLES | 0 | 0% |
| MID 3RD\_TACKLES | 0 | 0% |
| DEF 3RD\_TACKLES | 0 | 0% |
| TKLW\_TACKLES | 0 | 0% |
| TKL\_TACKLES | 0 | 0% |
| GCA\_GCA | 0 | 0% |
| SCA\_SCA | 0 | 0% |
| ONG\_TEAM.SUCCESS | 0 | 0% |
| ONGA\_TEAM.SUCCESS | 0 | 0% |
| PRGR\_RECEIVING | 0 | 0% |
| PPM\_TEAM.SUCCESS | 0 | 0% |
| MIN\_PLAYING | 0 | 0% |
| PRGP\_PROGRESSION | 0 | 0% |
| PRGC\_PROGRESSION | 0 | 0% |
| CRDR | 0 | 0% |
| CRDY | 0 | 0% |
| PKATT | 0 | 0% |
| PK | 0 | 0% |
| G\_MINUS\_PK | 0 | 0% |
| G+A | 0 | 0% |
| AST | 0 | 0% |
| GLS | 0 | 0% |
| STARTS\_PLAYING | 0 | 0% |
| SH\_STANDARD | 0 | 0% |
| MP\_PLAYING | 0 | 0% |
| MAX\_PLAYER\_VALUATION\_DATE | 0 | 0% |
| POSITION | 0 | 0% |
| AGE | 0 | 0% |
| CURRENT\_INTERNATIONAL | 0 | 0% |
| UrlFBref | 0 | 0% |
| LEAGUE\_COUNTRY | 0 | 0% |
| LEAGUE\_NAME | 0 | 0% |
| CLUB\_HREF | 0 | 0% |
| PLAYER\_HREF | 0 | 0% |
| PRGR\_PROGRESSION | 0 | 0% |
| SOT\_STANDARD | 0 | 0% |
| UNSUB\_SUBS | 0 | 0% |
| TOTDIST\_CARRIES | 0 | 0% |
| SUBS\_SUBS | 0 | 0% |
| COMPL\_STARTS | 0 | 0% |
| PLAYER\_VALUE | 0 | 0% |
| REC\_RECEIVING | 0 | 0% |
| DIS\_CARRIES | 0 | 0% |
| MIS\_CARRIES | 0 | 0% |
| CPA\_CARRIES | 0 | 0% |
| FINAL\_THIRD\_CARRIES | 0 | 0% |
| PRGC\_CARRIES | 0 | 0% |
| PRGDIST\_CARRIES | 0 | 0% |
| CARRIES\_CARRIES | 0 | 0% |
| FK\_STANDARD | 0 | 0% |
| TKLD\_TAKE | 0 | 0% |
| SUCC\_TAKE | 0 | 0% |
| ATT\_TAKE | 0 | 0% |
| LIVE\_TOUCHES | 0 | 0% |
| ATT PEN\_TOUCHES | 0 | 0% |
| ATT 3RD\_TOUCHES | 0 | 0% |
| MID 3RD\_TOUCHES | 0 | 0% |
| DEF 3RD\_TOUCHES | 0 | 0% |
| DEF PEN\_TOUCHES | 0 | 0% |
| TOUCHES\_TOUCHES | 0 | 0% |
| ERR | 0 | 0% |

Table : Player Market Value Bin

|  |  |  |
| --- | --- | --- |
| **PLAYER\_VALUE Bin** | **Player Count** | **Percentage of Players** |
| (0.0, 25000000.0] | 2141 | 89.81% |
| (25000000.0, 50000000.0] | 168 | 7.05% |
| (50000000.0, 75000000.0] | 47 | 1.97% |
| (75000000.0, 100000000.0] | 20 | 0.84% |
| (100000000.0, 125000000.0] | 5 | 0.21% |
| (125000000.0, 150000000.0] | 1 | 0.04% |
| (150000000.0, 175000000.0] | 0 | 0.00% |
| (175000000.0, 200000000.0] | 2 | 0.08% |

Table : Numerical features describe table

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **variable** | **count** | **mean** | **std** | **min** | **25%** | **50%** | **75%** | **max** |
| AGE | 2384 | 26.2 | 4.4 | 17 | 23 | 26 | 29 | 41 |
| MIN\_PLAYING | 2384 | 1328.4 | 948.4 | 1 | 475.25 | 1286 | 2108 | 3420 |
| DIST\_STANDARD | 2384 | 17.5 | 4.9 | 2.8 | 14.2 | 17.2 | 20.5 | 37.5 |
| DEF 3RD\_TOUCHES | 2384 | 233.5 | 273.0 | 0 | 35 | 127 | 341 | 1665 |
| ATT 3RD\_TOUCHES | 2384 | 206.7 | 204.5 | 0 | 43 | 147 | 314 | 1270 |
| ATT PEN\_TOUCHES | 2384 | 31.2 | 37.3 | 0 | 6 | 19 | 41 | 302 |
| ATT\_TAKE | 2384 | 27.0 | 31.4 | 0 | 5 | 16 | 38 | 306 |
| TOTDIST\_CARRIES | 2384 | 2648.4 | 2350.8 | 0 | 703.75 | 2149 | 3969.75 | 14430 |
| CPA\_CARRIES | 2384 | 6.3 | 10.9 | 0 | 0 | 2 | 8 | 140 |
| MIS\_CARRIES | 2384 | 21.7 | 21.3 | 0 | 5 | 16 | 31 | 139 |
| SUBS\_SUBS | 2384 | 6.4 | 5.4 | 0 | 2 | 5 | 9 | 29 |
| UNSUB\_SUBS | 2384 | 5.8 | 6.5 | 0 | 1 | 4 | 9 | 35 |
| PLUS\_PER\_\_MINUS\_\_TEAM.SUCCESS | 2384 | 0.0 | 12.7 | -46 | -6 | -1 | 5 | 58 |
| TOTDIST\_TOTAL | 2384 | 9168.7 | 9035.4 | 0 | 1951.8 | 6344 | 13902.8 | 54839 |
| CRSPA | 2384 | 3.0 | 4.8 | 0 | 0 | 1 | 4 | 37 |
| FLS | 2384 | 17.8 | 14.4 | 0 | 6 | 16 | 27 | 81 |
| FLD | 2384 | 16.8 | 16.2 | 0 | 4 | 13 | 24 | 122 |
| OFF | 2384 | 2.7 | 4.6 | 0 | 0 | 1 | 3 | 34 |
| CRS | 2384 | 25.6 | 39.9 | 0 | 2 | 9 | 33 | 393 |
| WON\_AERIAL | 2384 | 20.5 | 23.9 | 0 | 4 | 12 | 29 | 214 |
| LOST\_AERIAL | 2384 | 20.7 | 20.4 | 0 | 6 | 16 | 29 | 191 |
| ATT 3RD\_TACKLES | 2384 | 3.2 | 3.5 | 0 | 0 | 2 | 5 | 26 |
| LOST\_CHALLENGES | 2384 | 12.4 | 11.7 | 0 | 3 | 9 | 18 | 75 |
| PASS\_BLOCKS | 2384 | 11.9 | 10.2 | 0 | 3 | 10 | 18 | 52 |
| TKL+INT | 2384 | 37.9 | 33.8 | 0 | 9.75 | 29 | 59 | 195 |

Table : Numerical features describe table after Handling Outliers

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **variable** | **count** | **mean** | **std** | **min** | **25%** | **50%** | **75%** | **max** |
| AGE | 2384 | 26.1 | 4.4 | 17 | 23 | 26 | 29 | 39 |
| MIN\_PLAYING | 2384 | 1328.4 | 948.4 | 1 | 475.25 | 1286 | 2108 | 3420 |
| DIST\_STANDARD | 2384 | 17.4 | 4.8 | 2.8 | 14.2 | 17.2 | 20.4 | 32.1 |
| DEF 3RD\_TOUCHES | 2384 | 210.5 | 232.0 | 0 | 35 | 127 | 318 | 1052 |
| ATT 3RD\_TOUCHES | 2384 | 198.0 | 188.2 | 0 | 43 | 147 | 304.25 | 818 |
| ATT PEN\_TOUCHES | 2384 | 27.8 | 29.8 | 0 | 6 | 19 | 38 | 143 |
| ATT\_TAKE | 2384 | 24.3 | 25.7 | 0 | 5 | 16 | 35 | 121 |
| TOTDIST\_CARRIES | 2384 | 2548.3 | 2164.4 | 0 | 703.75 | 2149 | 3878.75 | 9699 |
| CPA\_CARRIES | 2384 | 4.9 | 7.1 | 0 | 0 | 2 | 7 | 38 |
| MIS\_CARRIES | 2384 | 20.5 | 18.9 | 0 | 5 | 16 | 30 | 85 |
| SUBS\_SUBS | 2384 | 6.2 | 5.1 | 0 | 2 | 5 | 9 | 22 |
| UNSUB\_SUBS | 2384 | 5.3 | 5.6 | 0 | 1 | 4 | 8 | 25 |
| PLUS\_PER\_\_MINUS\_\_TEAM.SUCCESS | 2384 | -0.4 | 11.4 | -38 | -6 | -1 | 4 | 38 |
| TOTDIST\_TOTAL | 2384 | 8695.0 | 8220.2 | 0 | 1951.8 | 6342 | 13346.8 | 35855 |
| CRSPA | 2384 | 2.5 | 3.6 | 0 | 0 | 1 | 4 | 17 |
| FLS | 2384 | 17.5 | 13.7 | 0 | 6 | 16 | 27 | 60 |
| FLD | 2384 | 15.9 | 14.3 | 0 | 4 | 13 | 24 | 65 |
| OFF | 2384 | 2.1 | 3.1 | 0 | 0 | 1 | 3 | 16 |
| CRS | 2384 | 21.0 | 28.4 | 0 | 2 | 9 | 30 | 145 |
| WON\_AERIAL | 2384 | 18.7 | 19.5 | 0 | 4 | 12 | 28 | 92 |
| LOST\_AERIAL | 2384 | 19.3 | 16.7 | 0 | 6 | 16 | 28 | 81 |
| ATT 3RD\_TACKLES | 2384 | 2.9 | 3.0 | 0 | 0 | 2 | 5 | 13 |
| LOST\_CHALLENGES | 2384 | 11.8 | 10.5 | 0 | 3 | 9 | 18 | 47 |
| PASS\_BLOCKS | 2384 | 11.6 | 9.7 | 0 | 3 | 10 | 18 | 42 |
| TKL+INT | 2384 | 36.9 | 32.1 | 0 | 9.75 | 29 | 58 | 139 |

**Figures**

Figure : Histogram of Player Market Value

A graph with numbers and lines

Description automatically generated

Figure : Box Plot of Player Market Value

A graph with numbers and lines

Description automatically generated with medium confidence

Figure : Histogram of Log Transformed Player Market Value

A graph of a market value

Description automatically generated

Figure : Bar Chart for Categorical Features

A screenshot of a graph

Description automatically generated

Figure : Histogram for Numerical Features

A chart of graphs and diagrams

Description automatically generated with medium confidence

A chart of graphs and diagrams

Description automatically generated with medium confidence

A chart of graphs and diagrams

Description automatically generated with medium confidence

Figure : Box Plot for Numerical Features

A group of blue and white diagrams

Description automatically generated with medium confidence

A group of blue and white diagrams

Description automatically generated with medium confidence

A group of blue and white diagrams

Description automatically generated with medium confidence

Figure : Heatmap for Numerical Subset 1

A blue and green squares

Description automatically generated

Figure : Heatmap for Numerical Subset 2

A blue and green squares

Description automatically generated

Figure : Heatmap for Numerical Subset 3

A blue and white squares

Description automatically generated

Figure : Heatmap for Numerical Subset 4

A blue and yellow squares

Description automatically generated

Figure : Heatmap for Numerical Subset 5

A blue and green squares with white text

Description automatically generated

Figure : Heatmap for Numerical Subset 6

A blue and yellow squares

Description automatically generated

Figure : Heatmap for Numerical Subset 7

A blue and green squares

Description automatically generated

Figure : Heatmap for Numerical Subset 8

A blue and green squares

Description automatically generated

Figure : Heatmap for Numerical Subset 9

A blue and green squares

Description automatically generated

# REFERENCES

Grus, J. (2015). Data science from scratch: first principles with Python. First edition. Sebastopol, CA, O'Reilly.

Knaflic, C. N. (2015). Storytelling with data: a data visualization guide for business professionals. Hoboken, New Jersey, John Wiley & Sons, Inc.

Provost, F., & Fawcett, T. (2013). Data science for business: [what you need to know about data mining and data-analytic thinking]. Sebastopol, Calif., O'Reilly.

Data Sources used:

<https://www.transfermarkt.com/>

<https://fbref.com/en/>

1. https://www.transfermarkt.com/ [↑](#footnote-ref-1)
2. https://fbref.com/en/ [↑](#footnote-ref-2)