NOTTINGHAM TRENT UNIVERSITY

SCHOOL OF SCIENCE AND TECHNOLOGY

Predictive Analytics in Sports

by

Soroush Ariana

in

2024

Project report in part fulfilment

of the requirements for the degree of

Master of Science

In

Software Engineering

I hereby declare that I am the sole author of this report. I authorize the Nottingham Trent University to lend this report to other institutions or individuals for the purpose of scholarly research.

I also authorize the Nottingham Trent University to reproduce this report by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.

Signature

Soroush Ariana

ABSTRACT

The accurate prediction of football players' market value is a critical concern for clubs and agents in the transfer market. Despite the availability of extensive player performance data, predicting market value remains challenging due to the complex interplay of numerous factors such as player statistics, age, and market conditions. This study addresses this problem by developing a desktop application that leverages machine learning models to estimate the market value of football players across different positions—midfielders, defenders, forwards, and goalkeepers.

The methodology involves the application of three regression models: Ordinary Least Squares, Decision Tree, and Random Forest. Each model was trained and evaluated using a dataset comprising player attributes and performance metrics relevant to their respective positions. The models were assessed based on their ability to predict the market value, with Mean Squared Error and Mean Absolute Error serving as the evaluation metrics. The desktop application was developed using Python to create a user-friendly interface that allows users to input player data and obtain market value predictions.

The findings demonstrate that the Random Forest model outperformed the other models in terms of prediction accuracy, particularly for players with more complex interactions between features. However, the study also highlights the limitations of the models, such as their sensitivity to outliers and the need for further refinement to enhance prediction reliability. The application provides a practical tool for stakeholders in the football industry, offering a data-driven approach to market value estimation, with the potential for future improvements and expansions.

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my supervisor, Dr. Ismahane Cheheb, for her invaluable guidance, support, and encouragement throughout the course of this project. Her insights and expertise have been instrumental in the successful completion of this work. I am sincerely thankful for her constant motivation and the knowledge she has shared with me during this journey.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

The data revolution has fundamentally transformed many sectors, including football, over the past decade. The European football market reached a size of 28.9 billion euros in 2018/19, marking a 32.57% growth over five years, despite the challenges posed by the COVID-19 pandemic. The transfer market is a particularly critical aspect, where player prices can reach astronomical levels. A notable example is Neymar Jr.'s record-breaking 222 million euro transfer from FC Barcelona to Paris Saint-Germain in 2017.

Significant risks characterize the transfer market, especially given the long-term contracts often involved, such as Saúl Ñíguez’s nine-year contract with Atletico Madrid or Lionel Messi's record-breaking four-year contract with FC Barcelona (Stępień, 2021). The application of quantitative approaches and big data in financial management could greatly benefit football teams, as illustrated by the 25-million-euro transfer of Giovani Lo Celso, who was later sold for 48 million euros.

The professional football transfer market is complex, with fixed-term contracts, various terms, bonuses, and adherence to regulations. Transfer fees can be substantial, with fourteen transactions historically surpassing the 100-million-euro mark (*The 100 most expensive football transfers of all time* | *Goal.com UK*, 2023). These complexities make it crucial for football teams to have accurate tools for financial decision-making.

Aims and objectives

This thesis seeks to address a fundamental question for football teams: What is the reasonable market value of a football player given their attributes? The primary aim is to develop a reliable and accurate system for predicting the market value of football players using machine learning techniques. The project is structured into several key phases:

1. **Data Selection and Pre-processing**:
   - Select and clean a suitable and reliable dataset to ensure high-quality data for machine learning tasks.
2. **Feature Engineering**:
   - Identify and design relevant features that significantly impact a player's market value, such as goals scored, assists, age, appearances, and injury records.
3. **Model Training and Evaluation**:
   - Train and evaluate multiple machine learning regression models to forecast player values, ensuring reliability and accuracy.
4. **Desktop Application Development**:
   - Develop a desktop application that integrates the best-performing model(s) to provide market value predictions for football players.
5. **Comprehensive Documentation**:
   - Document the entire process meticulously, producing a detailed report that includes the methodology, findings, and insights.

By achieving these objectives, this project aims to create a robust tool that can accurately predict football players' market values, offering valuable insights for clubs, agents, and analysts in the industry.

Workflow Overview

The following flowchart (Figure 1) provides a visual summary of the research and development process undertaken in this thesis, outlining the key steps and methodologies employed. This flowchart captures the workflow guiding the project from problem identification to application development and evaluation, ensuring a clear understanding of the approach taken to achieve the set objectives.

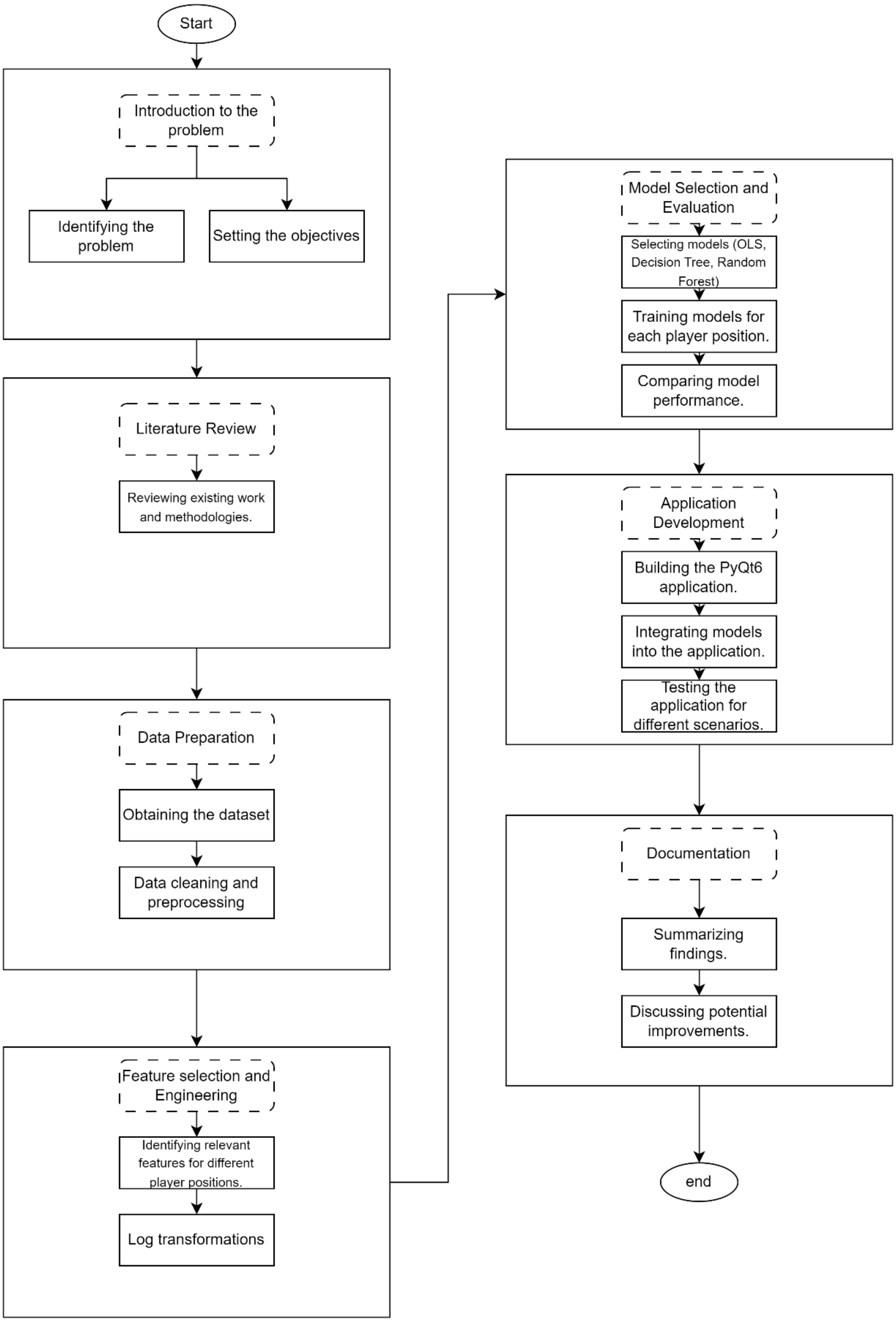

Start
Introduction to the problem
Identifying the problem
Setting the objectives
Literature Review
Reviewing existing work and methodologies.
Data Preparation
Obtaining the dataset
Data cleaning and preprocessing
Feature selection and Engineering
Identifying relevant features for different player positions.
Log transformations
Model Selection and Evaluation
Selecting models (OLS, Decision Tree, Random Forest)
Training models for each player position.
Comparing model performance.
Application Development
Building the PyQt6 application.
Integrating models into the application.
Testing the application for different scenarios.
Documentation
Summarizing findings.
Discussing potential improvements.
end

Figure 1: Flowchart representing the comprehensive workflow of the thesis.

This report provides a detailed account of the project focused on predicting football players' market values using machine learning. It begins with an **Introduction** that outlines the project's objectives and significance. Chapter 2, the **Literature Review**, explores key concepts, relevant studies, and identifies the research gap addressed by this project.

In Chapter 3, **New Ideas & Approach**, the report details the methodology, including the choice of regression models (OLS, Decision Tree, Random Forest), their development, and the features used for each player position. Chapter 4, **Implementation**, covers dataset handling, application development, and design considerations.

Chapter 5, **Evaluation**, assesses model performance and selects the best models for each position. Finally, Chapter 6, **Conclusions / Future Work**, summarizes the findings, discusses project limitations, and suggests directions for future research.

CHAPTER 2

LITERATURE REVIEW

Introduction

This chapter provides an overview of the key concepts, theories, and methodologies that form the foundation of this research. This chapter begins by defining the critical terms and concepts used in the study, such as the market value of football players, machine learning, regression models, and evaluation metrics. These definitions will set the stage for a deeper exploration of existing models and approaches for predicting football player market values, as well as the broader application of machine learning in sports analytics. The chapter concludes with an analysis of related work and the identification of gaps in the existing literature that this research seeks to address.

Definition and concepts

Market value

In the context of football, market value refers to the estimated worth of a player in the transfer market. This value is not fixed and can fluctuate based on various factors, including the player's performance, age, potential, contract length, and the overall demand for players with similar skill sets. Market value is often determined by clubs, agents, and analysts who consider a combination of subjective judgments and objective data, such as player statistics, market trends, and economic conditions.

Factors influencing a player's market value include:

- **Performance Metrics:** Goals scored, assists, defensive contributions, and overall match influence.
- **Player Attributes:** Age, physical fitness, injury history, and versatility.
- **Market Conditions:** Supply and demand for players in a specific position, economic conditions of football clubs, and transfer market trends.
- **External Factors:** Media coverage, sponsorship deals, and public perception.

Understanding the determinants of market value is crucial for accurate prediction, as it involves complex interactions between multiple variables.

Machine Learning

Machine learning (ML) is a subset of artificial intelligence (AI) that involves the development of algorithms capable of learning patterns from data and making predictions or decisions based on that data (Abramson, Braverman and Sebestyen, 1963; Mitchell, 1997). In predictive modeling, machine learning algorithms are trained on historical data to identify relationships between input features and output variables (Hastie, Tibshirani and Friedman, 2001). Once trained, these models can be used to predict outcomes for new, unseen data.

There are several types of machine learning, including:

- **Supervised Learning:** Where the model is trained on labeled data, meaning the input features and corresponding output labels are known. Regression models, such as those used in this study, fall under supervised learning (James *et al.*, 2023).

- **Unsupervised Learning:** Where the model is trained on unlabeled data and must identify patterns and structures without explicit instructions (Abramson, Braverman and Sebestyen, 1963).
- **Reinforcement Learning:** Where the model learns by interacting with an environment and receiving feedback in the form of rewards or penalties (Murphy, 2012).

In the context of this research, some of the machine learning-based known models are used to develop predictive models that estimate the market value of football players based on various performance and demographic features.

Related works

Player evaluation in various team sports, including baseball, basketball, and hockey, has been extensively studied, often focusing on rating systems like Elo and TrueSkill, which rank teams or players based on past performance and expected strength. However, these approaches typically do not account for specific match events or other quantitative performance factors. For example, Oliver Schulte and Zeyu Zhao's research on hockey player evaluation utilizes location-based data to assess NHL players, grouping them by similar play styles and impacts to better evaluate performance (Schulte *et al.*, 2017). This approach emphasizes the use of spatial data to capture the context of players' actions, thus offering novel insights through advanced clustering techniques. Despite these advantages, the focus on NHL players and the reliance on detailed location data may limit the applicability to other sports and contexts.

Studies like those by Decroos et al. have emphasized the importance of contextualizing player actions within the flow of a match (Decroos *et al.*, 2020). Their work introduces the VAEP (Valuing Actions by Estimating Probabilities) framework, which attempts to assign value to every action a player makes, such as passes or shots, by estimating the impact on a team's chances of scoring or conceding a goal. This method provides a more nuanced understanding of player performance beyond basic statistics like goals and assists, but it is complex and requires detailed event data, making it less accessible for clubs with limited resources.

In an effort to develop more dynamic and probabilistic models, Herbrich, et al. introduced TrueSkillTM, a Bayesian skill ranking system that extends the Elo system (Schölkopf, Platt and Hofmann, 2007). TrueSkill is well-regarded for its application in sports and gaming, providing a robust method for ranking players based on opponent strength and match results. However, as Lasek, Szlávik, and Bhulai noted in their research on ranking systems in association football, while these systems offer a reasonable estimate of player or team strength, they are limited by their reliance on past results and fail to incorporate comprehensive match event data, which can significantly impact performance assessments (Lasek, Szlávik and Bhulai, 2013).

Moving closer to the specific domain of soccer, Pappalardo and Cintia (2018) emphasized the importance of detailed match data in evaluating player performance (Pappalardo and Cintia, 2018). Their work highlighted the necessity of considering both individual player actions and overall team performance to accurately assess player contributions and value. This approach underscores the potential for a more granular analysis of soccer performance metrics to provide realistic evaluations of player impact.

Building on these foundations, Pappalardo et al. proposed a new paradigm for the objective rating of soccer players using machine learning (Pappalardo *et al.*, 2019). By creating feature vectors that assess player performance across multiple games and seasons, their data-driven approach offers advantages such as replicability, transparency, and the ability to customize evaluations for different leagues and events. Despite these benefits, the model's reliance on high-quality input data and the complexity of machine learning models present challenges, particularly for stakeholders like scouts and coaches who may struggle with understanding the outputs. Additionally, the system must contend with the dynamic nature of player performance, necessitating continuous data updates and model retraining.

Further refining the application of machine learning to soccer, Behravan and Razavi (2021) developed a method for estimating football players' market prices by integrating personal traits, player statistics, and past transfer information into a regression model (Behravan and Razavi, 2021). This approach aims to deliver more accurate and dynamic valuations compared to traditional methods. However, like other models, it faces limitations related to data quality, the interpretability of machine learning algorithms, and the need for continuous updates to reflect changing player values.

Yiğit, Samak, and Kaya contributed significantly to this field by developing a model that provides objective, statistically-based insights into player values, incorporating a broad range of factors, including performance indicators and physical characteristics (Yiğit, Samak and Kaya, 2020). Their use of advanced algorithms like Random Forest and XGBoost has demonstrated high prediction accuracy, though challenges remain in terms of data quality, model interpretability, and the resource-intensive nature of these models.

Addressing the "black-box" nature of traditional machine learning models, Huang and Zhang explored the application of Explainable AI (XAI) techniques to soccer player valuation (Huang and Zhang, 2023). By using SHAP values and partial dependence graphs, their work improves the transparency and interpretability of the valuation process, making it more accessible to stakeholders such as managers, coaches, and scouts. However, the reliance on sophisticated XAI methods introduces additional computational overhead and requires substantial expertise to implement effectively.

Bialkowski et al. investigated the integration of GPS tracking data with event-based match data to evaluate player performance (Bialkowski *et al.*, 2014). Their work focused on using physical metrics such as distance covered, speed, and acceleration in combination with tactical data to gain a deeper understanding of player roles and effectiveness in various phases of the game. This approach offers a more holistic view of player performance but also introduces challenges in data collection, processing, and interpretation.

Gudmundsson and Horton are also explored the use of deep learning techniques in predicting outcomes of football matches (Gudmundsson and Horton, 2018). Their study utilized recurrent neural networks (RNNs) to capture temporal dependencies in match sequences, which allowed for more accurate predictions of match outcomes compared to traditional statistical models. However, the requirement for extensive computational resources and expertise in deep learning may limit the practical application of these methods in everyday sports analytics.

Finally, a study by Elahi, Pandey, and Malhi offers an exhaustive comparison of various machine learning techniques for predicting football player market values (Elahi, Pandey and Malhi, 2023). Their research highlights the impact of feature selection and data preprocessing on model performance and seeks to identify the most effective method for practical use in the football transfer market. While their comparison provides valuable insights, the study's focus on past data may limit its ability to capture unforeseen market fluctuations or emerging trends. Additionally, the interpretability of the models remains an area requiring further attention to ensure that stakeholders understand the reasoning behind predictions.

Research Gap

Despite significant advancements in the application of machine learning and statistical models to player evaluation and market value estimation, several limitations and challenges remain unaddressed in the current literature.

1. **Lack of Contextual Integration**: While many studies, such as those by Pappalardo et al. and Behravan and Razavi (Pappalardo *et al.*, 2019; Behravan and Razavi, 2021), have made strides in incorporating player performance data and personal traits into valuation models, these approaches often fail to fully account for the dynamic and contextual factors that influence player performance and market value. Elements such as team dynamics, tactical roles, and individual player morale are rarely integrated into these models, leading to valuations that may not fully reflect real-world conditions.
2. **Model Interpretability and Complexity**: The complexity of advanced machine learning models, highlighted in works by Yiğit et al. and Huang and Zhang (Yiğit, Samak and Kaya, 2020; Huang and Zhang, 2023), presents a significant barrier to their adoption by stakeholders such as coaches, scouts, and managers. These models, while powerful, often operate as "black boxes," making it difficult for users to understand the reasoning behind predictions. Although Explainable AI (XAI) techniques offer some solutions, they add computational overhead and require expertise that may not be readily available.
3. **Data Quality and Availability**: A recurring issue across multiple studies is the heavy reliance on high-quality, comprehensive datasets. As noted in works by Schulte et al. and Elahi, Pandey, and Malhi (Schulte *et al.*, 2017; Elahi, Pandey and Malhi, 2023), the accuracy and reliability of machine learning models are contingent on the availability of robust data. However,

data collection in sports can be inconsistent, leading to potential biases and incomplete analyses.

4. **Over-reliance on Historical Data**: Many existing models, including those developed by Lasek et al. and Pappalardo and Cintia (Lasek, Szlávik and Bhulai, 2013; Pappalardo and Cintia, 2018), depend on historical data to predict future player values. This approach can be limiting, as it may not adequately capture emerging trends, unexpected market shifts, or novel player attributes that could influence future performance.
5. **Scalability and Resource Intensity**: The application of machine learning in sports analytics, as demonstrated by Yiğit et al. (Yiğit, Samak and Kaya, 2020) and others, often requires significant computational resources and expertise in data science. This scalability issue can hinder the practical application of these models, particularly in resource-constrained environments.

Addressing the Gap

Our research addresses the identified gaps by developing a more comprehensive and interpretable machine learning model for football player market value estimation. To achieve this, we integrate contextual factors such as team dynamics and tactical roles, providing a more holistic evaluation of player performance that goes beyond what current models offer.

Moreover, our work emphasizes the importance of robust feature engineering and data preprocessing to ensure the model's reliability, even when faced with data limitations. By incorporating mechanisms for continuous learning and adaptation, our model can respond to emerging trends and shifts in the football market, making it a dynamic and responsive tool for player valuation.

To operationalize these advancements, we have also developed a desktop application that integrates the best-performing machine learning models. This application allows users to input relevant player data and receive real-time market value predictions, thereby bridging the gap between complex model outputs and practical, user-friendly solutions.

In summary, our research fills the gaps identified in the literature by offering a more contextually aware, interpretable, and scalable solution for football player market value estimation, ultimately contributing to more informed and strategic decision-making in the football transfer market.

CHAPTER 3

NEW IDEAS & APPROACH

Introduction

Accurately predicting the market value of football players is a multifaceted challenge in the rapidly evolving domain of football analytics. While traditional methods like linear models and expert assessments have provided a foundation for player valuation, they often fall short in capturing the dynamic and contextual factors that significantly influence market value. These approaches may overlook the complex interactions between player attributes, team dynamics, tactical roles, and emerging trends in the football market.

This chapter introduces a novel approach that addresses these gaps by leveraging advanced machine learning techniques to enhance both the accuracy and interpretability of market value predictions. By integrating contextual factors such as team dynamics, player morale, and tactical roles, our model offers a more holistic and realistic representation of player performance. This integration not only improves the precision of market value predictions but also ensures that the insights are transparent and understandable to stakeholders like coaches, scouts, and managers.

To make this advanced model accessible and practical, we have developed a desktop application using the PyQt6 framework, detailed in Chapter 4. This application allows football scouts, analysts, and enthusiasts to easily input relevant player features and obtain accurate market value predictions based on the best-performing machine learning models for each position. The user-friendly interface ensures that the complexity of the underlying models does not hinder usability, making advanced analytics accessible to non-expert users.

Additionally, our approach emphasizes robust feature engineering and data preprocessing to address issues related to data quality and availability, ensuring the model's resilience in varying data conditions. Continuous learning mechanisms further enable the model to adapt to emerging trends and market shifts, keeping predictions relevant and timely.

In summary, this chapter presents an innovative methodology for football player market value estimation and a practical desktop application that embodies this approach. By overcoming the limitations of existing models and offering a dynamic, context-aware, and scalable solution, this work aims to support more informed and strategic decision-making within the football transfer market.

Proposed Methodology

Methodology

The development of the football player market value prediction model and its integration into a desktop application were guided by two complementary methodologies: the **CRISP-DM (Cross Industry Standard Process for Data Mining)** framework for the machine learning aspect, and the **Waterfall** methodology from the Software Development Life Cycle (SDLC) for the software engineering part.

CRISP-DM, or Cross Industry Standard Process for Data Mining, is a robust and flexible methodology specifically designed for data mining and machine learning projects (Shearer, 2000). Developed in the late 1990s by industry leaders, CRISP-DM has become the de facto standard for data-driven projects across various domains due to its comprehensive and structured approach. The methodology is particularly effective in scenarios where the project involves iterative cycles of data exploration, modelling, and validation. It consists of six phases: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment. CRISP-DM was chosen for this project because it provides a structured yet flexible approach that supports iterative refinement of both the data and the models. This adaptability is crucial when dealing with the complexities and unpredictabilities inherent in machine learning projects, such as the one undertaken for predicting football player market values.

The Waterfall model is one of the earliest and most straightforward methodologies within the Software Development Life Cycle (SDLC). Originating from traditional engineering practices, it follows a linear and sequential approach to software development (Royce, 1970). The Waterfall model is characterized by its clear structure, where each phase of the project must be completed before the next one begins, ensuring a well-defined pathway from project initiation to completion. It involves distinct stages, including Requirement Analysis, System Design, Implementation, Integration and Testing, Deployment, and Maintenance. This methodology was selected for the software engineering component of this project due to its structured and predictable nature. Given the well-defined requirements of the desktop application, the Waterfall model provided a clear pathway for development, ensuring that each phase was thoroughly completed before moving on to the next. This approach minimized risks and facilitated a stable and reliable final product, which is crucial for the practical deployment of the predictive models in a real-world setting.

Application of CRISP-DM Framework

1. **Business Understanding**:

    - **Objective**: The primary goal was to develop a predictive model that accurately estimates the market value of football players based on various attributes and to make this model accessible through a user-friendly desktop application.

- **Challenges**: Key challenges included identifying relevant features, addressing data quality issues, and selecting appropriate machine learning models.

2. **Data Understanding**:

   - **Dataset Exploration**: A pre-existing dataset published by (Stępień, 2021), containing player statistics and market values, was utilized. The dataset was examined to understand its structure, identify missing values, and assess its suitability for the modeling task.

   - **Feature Selection**: Important features relevant to each player position (goalkeeper, defender, midfielder, forward) were selected based on domain knowledge and exploratory data analysis.

3. **Data Preparation**:

   - **Preprocessing**: This phase involved handling missing values, performing log transformations to normalize the data, and filtering data by player position.

   - **Feature Engineering**: Additional features were engineered to enhance the model’s predictive power, incorporating contextual factors that influence player performance and value.

4. **Modeling**:

   - **Model Selection**: The project experimented with multiple machine learning models, including Ordinary Least Squares (OLS), Decision

Tree, and Random Forest, to determine the best-performing model for each player position.

- **Training and Evaluation**: Each model was trained and evaluated using metrics like Mean Squared Error and Mean Absolute Error. The best model for each position was selected based on its performance on the validation set.

Application of Waterfall SDLC Methodology

1. **Requirement Analysis**:
    - Detailed requirements for the desktop application were gathered, focusing on user needs such as ease of use, model integration, and result visualization.
2. **System Design**:
    - Based on the requirements, the architecture of the desktop application was designed. This included the user interface layout using the PyQt6 framework and the integration points for the predictive models.
3. **Implementation**:
    - The desktop application was developed in a linear fashion. The user interface was created first, followed by the integration of the machine learning models. Each step was completed before moving on to the next, ensuring a stable and reliable build.
4. **Integration and Testing**:
    - After development, the application was rigorously tested to ensure that all components functioned as expected. The predictive models

were tested for accuracy, and the user interface was evaluated for usability.

5. **Deployment**:
    - The application was deployed for use by football scouts and analysts. This phase also included final adjustments based on user feedback.
6. **Maintenance**:
    - Although maintenance was not a major focus, provisions were made to ensure the application could be updated as needed, especially as new data or requirements emerged.

By combining the CRISP-DM framework for machine learning with the Waterfall SDLC methodology for software development, this project ensured a thorough and systematic approach to both the predictive modeling and the development of a user-friendly desktop application. The integration of these methodologies allowed for a well-rounded solution that balances technical accuracy with practical usability.

The following Gantt chart (Figure 2) outlines the timeline for the key activities. This chart provides a visual representation of the project's schedule, highlighting the planning and execution phases, and ensuring that all tasks are completed within the allocated timeframes.

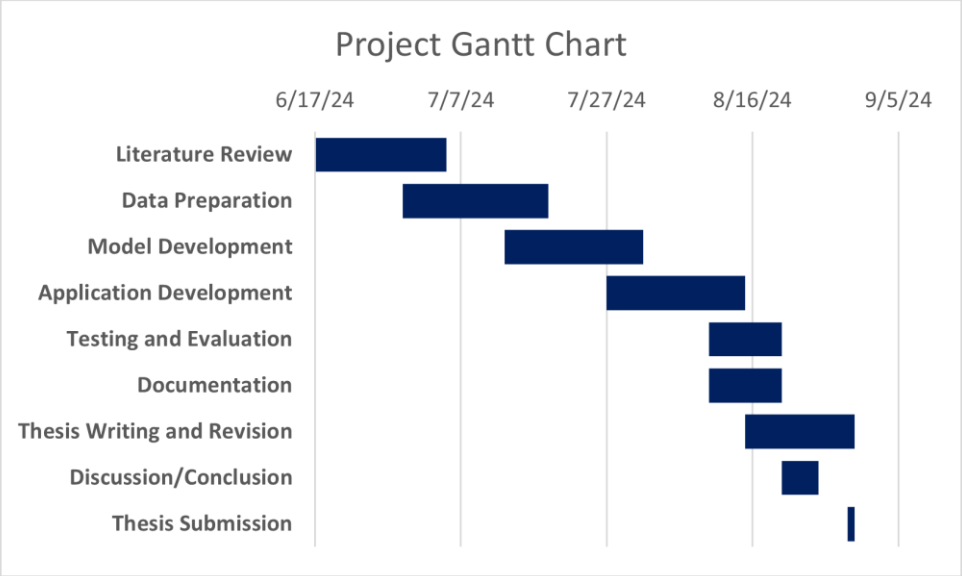


Figure 2: Project Gantt chart.

With a structured methodology in place, the next step involved selecting and implementing the most suitable machine learning models for predicting football player market values. The following sections provide a brief overview of the three key models employed in this project: Ordinary Least Squares (OLS), Decision Tree, and Random Forest.

Ordinary Least Squares (OLS) Regression

Ordinary Least Squares (OLS) regression is a linear regression model that estimates the relationship between the independent variables (features) and the dependent variable (target) by minimizing the sum of the squared differences between the observed and predicted values (Hastie, Tibshirani and Friedman, 2001). The OLS model is expressed by the following formula:

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

where:

- $\hat{y}$ is the predicted value,
- β_0 is the intercept,

- $\beta_1, \beta_2, \cdots, \beta_p$ are the coefficients of the independent variables $x_1, x_2, \cdots, x_p$.

The coefficients β are determined by minimizing the cost function, which is the sum of squared residuals (the differences between observed and predicted values):

Cost Function (MSE) $= \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$

where:

- n is the number of observations,
- y_i is the observed value,
- $\hat{y}_i$ is the predicted value.

Decision Tree Regressor

A Decision Tree Regressor is a non-linear model that predicts the value of a target variable by learning simple decision rules inferred from the data features (Breiman, 2017). It recursively partitions the data space into smaller regions and fits a simple prediction model (usually the mean of the target values) within each region.

The decision tree splits the data based on the feature that minimizes the mean squared error at each node. The prediction for a new observation is made by following the decision rules from the root to a leaf node, where the final prediction is the average of the target values in that node.

The splitting criterion at each node is defined as:

$$\text{Split Criterion } = \operatorname{argmin}_{\theta} \sum_{j=1}^{J} \sum_{i \in N_j} \left(y_i - \hat{y}_{N_j}\right)^2$$

where:

- N_j represents the samples in the jth node,
- $\hat{y}_{N_j}$ is the mean of the target values in node N_j.

Random Forest Regressor

The Random Forest Regressor is an ensemble learning method that builds multiple decision trees and merges their predictions to produce a more accurate and robust model (Breiman, 1996). Each tree in the forest is trained on a different random subset of the data (using bootstrapping) and a random subset of features. The final prediction is obtained by averaging the predictions of all the trees.

The formula for the prediction in a Random Forest model is:

$$\hat{y}_{\mathrm{RF}} = \frac{1}{T}\sum_{t=1}^{T} \hat{y}^{(t)}$$

where:

- T is the number of trees in the forest,
- $\hat{y}^{(t)}$ is the prediction from the tth tree.

By aggregating multiple decision trees, Random Forest reduces the variance of the model and improves its generalization ability, making it a powerful tool for complex regression tasks.

Justification for the Selected Models

In this study, three different regression models—Ordinary Least Squares, Decision Tree, and Random Forest—were selected to predict the market value of football players across various positions. The choice of these models was driven by their distinct characteristics and the diverse insights they offer in the context of regression tasks.

Ordinary Least Squares (OLS)

The OLS regression model was chosen for its simplicity and interpretability. OLS is a linear regression method that estimates the relationship between the independent variables (features) and the dependent variable (market value) by minimizing the sum of squared differences between the observed and predicted values. This model is particularly useful for understanding the direct impact of each feature on the market value, making it an excellent choice for positions like goalkeepers, where fewer, more linear relationships might dominate.

Key reasons for choosing OLS:

- **Interpretability**: OLS provides clear insights into the relationship between each feature and the target variable, which is valuable for understanding how specific player attributes influence their market value.

- **Efficiency**: OLS is computationally efficient, making it a suitable choice when a quick, initial model is needed for comparison with more complex methods.

- **Baseline Performance**: As a simple linear model, OLS serves as a benchmark to evaluate the performance of more complex, non-linear models.

Decision Tree Regressor

The Decision Tree model was selected for its ability to handle non-linear relationships and interactions between features. Unlike OLS, which assumes a linear relationship, the Decision Tree algorithm splits the data into smaller, homogeneous groups, making predictions based on the mean value of the target variable in each group. This model is particularly effective for positions where non-linear patterns exist, such as midfielders, where a variety of skills and interactions between attributes are crucial.

Key reasons for choosing Decision Tree:

- **Non-Linear Relationships**: Decision Trees can capture complex interactions and non-linear relationships between features, which is important for modeling player values in positions where such patterns are expected.
- **Interpretability**: Despite being non-linear, Decision Trees are still interpretable, as the splits in the tree can be traced back to understand the decision-making process.
- **Handling Mixed Data**: Decision Trees can handle both numerical and categorical data effectively, making them versatile for the diverse range of features in the dataset.

Random Forest Regressor

The Random Forest model was chosen as an advanced ensemble method that builds upon the strengths of Decision Trees. By creating an ensemble of multiple Decision Trees and averaging their predictions, Random Forest mitigates the overfitting tendency of individual trees and enhances predictive accuracy. This model is particularly useful for predicting market values in positions like defenders and forwards, where complex, non-linear interactions and a high variance in features are expected.

Key reasons for choosing Random Forest:

- **High Predictive Accuracy**: Random Forest typically outperforms individual Decision Trees by reducing overfitting and capturing a broader range of patterns in the data.

- **Robustness**: The ensemble nature of Random Forest makes it more robust to noise and variations in the data, leading to more reliable predictions.

- **Feature Importance**: Random Forest provides a measure of feature importance, offering insights into which player attributes are most influential in determining market value.

Complementary Strengths of the Models

The combination of these three models allows for a comprehensive analysis of the factors influencing football player market values across different positions. By leveraging the strengths of linear (OLS) and non-linear (Decision Tree, Random Forest) methods, the study can account for both simple and complex relationships within the data, ensuring that the best model is chosen for each player position based on empirical performance.

Creation of models

In this section, we explain the implementation and creation of the predictive models. Before discussing the models, themselves, it's essential to outline the data preparation process.

After preprocessing the dataset and consolidating data from all seasons into a single file, the dataset was organized and split based on player positions:

1. **Goalkeepers**: The dataset was filtered to include only players listed as 'Goalkeeper'. One player, Emiliano Viviano, was excluded due to issues with his data.

2. **Defenders**: Players categorized as 'Defender' were extracted based on the position label.

3. **Forwards**: The dataset was divided into two subsets for players labeled as 'Forward' and 'attack'. These subsets were then combined to form a comprehensive dataset for forwards.

4. **Midfielders**: Players labeled as 'Midfielder' and 'midfield' were similarly extracted and combined to create the dataset for midfielders.

This approach ensures that each dataset is tailored to the specific position, facilitating more accurate model training and evaluation. To provide an overview of the dataset distribution, a table (table 2) summarizing the number of rows corresponding to each player position is included.

Table 1: Table displaying the count of observations utilized in modelling for each position.

Position	Number of observations
Goalkeepers	535
Defenders	2375
Midfielders	1384
Forwards	1379

To predict the target variable for each player position, we developed three predictive models: Ordinary Least Squares (OLS), Decision Tree Regressor, and Random Forest Regressor. Each model was designed and implemented following specific steps to ensure accurate and reliable predictions. The OLS model uses a linear approach, where the dataset is split into training (80%) and test (20%) sets. The target variable, 'value,' is log-transformed to 'ln_value' to address potential issues with variance and skewness, improving model performance and stability. The model is then trained using a formula that integrates 'ln_value' with the selected features, estimating the relationship between the target and predictors to model linear relationships effectively. The model’s performance is evaluated by predicting 'ln_value' on the test set, using performance metrics to assess its accuracy.

The Decision Tree Regressor captures complex, non-linear relationships between the features and the target variable. The dataset is similarly split, and the target variable is log-transformed to normalize the data and enhance predictive capability. A Decision Tree Regressor is then trained on the log-transformed target variable and selected features, constructing a tree-like structure where each node represents a decision rule. This model is evaluated on the test set to determine its performance in handling non-linear relationships and feature interactions.

Finally, the Random Forest Regressor improves prediction accuracy by aggregating the results of multiple decision trees. After splitting the dataset and applying the same log transformation to the target variable, a Random Forest Regressor is trained on 'ln_value' and the features. This ensemble approach enhances the model’s robustness by averaging predictions from multiple trees, reducing overfitting, and improving generalization. The model's predictions are evaluated on the test set to gauge its performance in aggregating the results from multiple decision trees.

Feature set for each position

In football analytics, the selection and extraction of features play a critical role in the development and accuracy of predictive models. Each football position on the field has distinct responsibilities, skills, and contributions to the game, which necessitates the use of tailored feature sets to effectively model performance.

Each football position has unique roles and responsibilities that influence performance metrics. The selected features are designed to capture the nuances of these roles. Midfielders act as the link between defense and attack, and their performance is often evaluated based on their ability to control the game, distribute the ball, and contribute both defensively and offensively. Features like passing accuracy, carries, and tackles are crucial for assessing their overall impact on the game. By focusing on metrics such as passes into the final third and xG/xA per 90, we can better understand their contributions to offensive play and ball progression.

The primary role of defenders is to prevent the opposition from scoring and to manage threats in and around the penalty area. Features such as aerial duels won, touches in defensive penalty areas, and xGA are essential to evaluate their effectiveness in blocking, intercepting, and managing offensive threats. Metrics like passes ground also help in assessing their ability to transition play from defense to attack.

Forwards are primarily responsible for scoring goals and creating scoring opportunities. Metrics such as goals scored, dribbles completed, and xG are critical for evaluating their offensive output and ability to break down defenses. Features like GCA (Goals Created) and touches in the attacking penalty area are important to measure their role in setting up and converting chances.

Goalkeepers are tasked with preventing goals and managing shots on target. Features such as clean sheets, saves per shot on target, and passes launched are tailored to assess their shot-stopping abilities, distribution skills, and overall effectiveness in maintaining a strong defensive record.

Using position-specific features allows the models to focus on the most relevant aspects of performance for each role, thereby enhancing prediction accuracy. For example, Midfielders benefit from features that capture both defensive and offensive contributions, allowing the model to accurately predict their overall impact on the game. Metrics like carry distance and tackles won provide insight into their defensive capabilities, while passes into the final third and xG reflect their offensive contributions. In other side, defenders require features that capture their defensive effectiveness and contributions to the team’s overall defensive stability. By including metrics like aerials won (%) and touches in defensive areas, the model can better predict a defender’s ability to handle various defensive situations. Forwards also need features that capture their scoring potential and playmaking abilities. Metrics like goals scored and dribbles completed are essential for assessing their offensive impact, while GCA and xG provide insight into their ability to contribute to goal-scoring opportunities. Finally, Goalkeepers benefit from features that measure their shot-stopping effectiveness and distribution skills. By focusing on metrics like clean sheets and saves per shot on target, the model can more accurately evaluate a goalkeeper’s performance in preventing goals and managing the defense.

The selection of specific feature sets for each football position is a crucial step in developing effective predictive models. Football datasets often contain a vast number of features, many of which may not be relevant to every position. By carefully choosing a specific set of features for each position, we can reduce dimensionality, which allows us to concentrate on the most impactful features for each role. This simplification not only enhances the interpretability of the model but also improves its overall effectiveness.

Moreover, using a targeted set of features enhances the model training process. With fewer, more relevant features, the model can be trained more efficiently, reducing the risk of overfitting and improving its ability to generalize to new data. This careful selection of features ensures that the model is not overwhelmed by irrelevant data, allowing it to focus on what truly matters for each position.

In addition to feature selection, capturing relevant performance metrics for different positions is essential to accurately assess the contributions of players. For example, midfielders require metrics related to ball distribution, playmaking, and defensive actions, reflecting their central role in both creating and disrupting play. For defenders, it is crucial to focus on metrics that capture defensive actions, aerial duels, and their impact on preventing goals, as these are key indicators of their defensive prowess. Forwards, on the other hand, are best evaluated using offensive metrics such as goal-scoring and chance creation, which highlight their effectiveness in contributing to the team's attacking efforts. Goalkeepers require metrics related to shot-stopping, clean sheets, and distribution, which are essential for assessing their performance in goalkeeping duties.

By extracting and focusing on these position-specific feature sets, the predictive models are better aligned with the unique responsibilities and contributions of each football position. This approach leads to more accurate and meaningful predictions, as the models are tailored to capture the nuances of each role on the field.

In this section, we provide a comprehensive overview of the features used to predict player performance across different football positions. The features selected for each position are drawn from a common set, with specific features tailored to the unique demands of each role. Below is a table summarizing the features and their descriptions, followed by a list of the features applied to each position.

Table 2: Detailed descriptions of the features used to predict player performance across various football positions. These features are selected for their relevance to each role, encompassing both common attributes shared among positions and specific metrics tailored to the demands of midfielders, defenders, forwards, and goalkeepers.

Feature	**Description**
Age	A fundamental feature that correlates with a player's experience, physical development, and decision-making skills.
CL	Reflects a player's experience in high-pressure and high-stakes matches, particularly in the UEFA Champions League.
Goals	Measures the number of goals scored by the player, indicating their offensive contributions and ability to influence the game's outcome.
Passes Completed (Short)	Represents the accuracy of short-range passes, crucial for maintaining possession and facilitating team play, especially for midfielders.
Passes into Final Third	Assesses the player's ability to progress the ball into the opponent's final third, essential for creating scoring opportunities.
Pts	Aggregates various performance metrics to reflect the player's overall contribution to the team's success.
xG	Quantifies the quality of scoring opportunities created or participated in by the player
xGA	Measures the quality of scoring chances conceded while the player is on the field, reflecting their defensive contributions.
xG/xA per 90	A combined metric of Expected Goals and Expected Assists per 90 minutes, evaluating the player's overall offensive output.
Carry Distance	Highlights the distance a player covers while carrying the ball, indicating dribbling ability and the capacity to drive play forward.
Tackles Won	Measures the player's effectiveness in defensive situations, specifically their ability to win tackles and recover possession.
Passes Ground	Reflects the accuracy of ground passes, important for defenders who initiate attacks from the back.
Touches in Attacking Penalty Area	Indicates the player's involvement in key scoring areas, particularly for forwards, reflecting their positioning and ability to capitalize on opportunities.

Touches in Defensive Penalty Area	Indicates a defender's involvement in critical defensive actions within the penalty area, crucial for managing threats.
Aerials Won	The percentage of aerial duels won by the player, reflecting their effectiveness in dealing with high balls, especially important for defenders.
GCA	Measures the player's ability to create goal-scoring opportunities for teammates, highlighting their playmaking skills.
Dribbles Completed	Represents the player's success in beating defenders through dribbling, contributing to offensive play.
Wins (Gk)	The number of matches won by the team when the goalkeeper is on the field, indicating their contribution to positive results.
Draws (Gk)	The number of matches that ended in draws, reflecting the goalkeeper's ability to prevent losses.
Passes % Launched (Gk)	Measures the accuracy of long passes made by the goalkeeper, important for distributing the ball and initiating counter-attacks.
psnpxg per Shot on Target Against	Expected Goals Saved per shot on target, assessing the goalkeeper's shot-stopping abilities and defensive reliability.
Clean Sheets	The number of games in which the goalkeeper did not concede any goals, reflecting their effectiveness in maintaining a strong defensive record.
League Dummies	Categorical variables capturing the league (Premier League, Ligue 1, Serie A, La Liga, Bundesliga) in which the player competes, influencing performance metrics.

Feature Sets for Each Position

- **Midfielders:**
    - age, goals, CL, passes_completed_short, passes_into_final_third, Pts, xG, xGA, xg_xa_per90, carry_distance, tackles_won, isPremierLeague, isLigue1, isSerieA, isLaLiga, isBundesliga

- **Defenders:**
  - age, CL, goals, xg_xa_per90, passes_ground, touches_att_pen_area, touches_def_pen_area, aerials_won_pct, Pts, xGA, xG, isPremierLeague, isLigue1, isSerieA, isLaLiga, isBundesliga
- **Forwards:**
  - age, CL, goals, gca, Pts, xG, xGA, dribbles_completed, xg_xa_per90, touches_att_pen_area, passes_into_final_third, isPremierLeague, isLigue1, isSerieA, isLaLiga, isBundesliga
- **Goalkeepers:**
  - age, CL, wins_gk, draws_gk, passes_pct_launched_gk, psnpxg_per_shot_on_target_against,clean_sheets,isPremierLeague, isLigue1, isSerieA, isLaLiga, isBundesliga

Professional, Social, Ethical, and Legal Issues

The completion of this project involves several professional, social, ethical, and legal considerations, specific to the domain of football analytics and machine learning.

Professional Issues

The development of this application underscores the critical importance of accurate data analysis and model selection in professional football analytics. Given that the application is designed to predict football player market values, which could influence decisions related to player transfers, valuations, and team strategies, it is imperative that the models used are both robust and reliable. A specific challenge faced was ensuring that the models selected (e.g., OLS, Decision Tree, Random Forest) provided accurate predictions across different player positions. To address this, a rigorous model evaluation process was implemented, involving cross-validation and the use of multiple metrics (MSE, MAE) to ensure the models' performance met professional standards. Additionally, transparency was enhanced by documenting the model selection process and providing clear explanations for the predictions, which helps in building trust among users like scouts and analysts.

Social Issues

The application could have significant social implications within the football industry, particularly in how player valuations might affect contract negotiations, team compositions, and public perceptions. In high-stakes environments like football, reliance on algorithmic predictions can lead to unintended social consequences, such as undervaluing certain players based on incomplete or biased data. To mitigate this, the project incorporated extensive feature engineering and data preprocessing to ensure that the models accounted for a wide range of relevant factors, thus providing a more comprehensive and fair valuation. Additionally, by creating a user-friendly desktop application, the project ensures that a broader range of stakeholders, including those without technical expertise, can effectively utilize the tool, promoting inclusivity in decision-making processes.

Ethical Issues

Ethically, the use of machine learning for predicting player market values presents challenges related to bias and fairness. There was a risk that the models could reinforce existing biases in the data, potentially leading to discriminatory valuations based on age, nationality, or other non-performance-related factors. To address these ethical concerns, the project carefully selected and engineered features that focus on performance-related attributes, minimizing the influence of potentially biased factors. Furthermore, during the data preprocessing stage, checks were implemented to detect and mitigate biases, ensuring that the model predictions are fair and equitable across different player demographics.

Legal Issues

Legally, the project adhered to data protection laws and intellectual property rights. Since the dataset used is publicly available, there were fewer concerns about data ownership. However, ensuring the confidentiality and security of the data during processing was still a priority. The project also respected the intellectual property rights associated with the machine learning models and tools used, ensuring all components were either open-source or appropriately licensed.

CHAPTER 4

IMPLEMENTATION

Introduction

This chapter outlines the implementation process of the study, focusing on the dataset used, the programming languages employed, and the libraries integrated into the development of the predictive models. The implementation phase is crucial as it bridges the theoretical underpinnings discussed in earlier chapters with the practical aspects of model development and evaluation.

The chapter begins with a detailed description of the dataset utilized in this research. This section covers the dataset's structure, and the preprocessing steps undertaken to prepare the data for analysis. Understanding the dataset's characteristics and any associated issues is essential for comprehending the context in which the predictive models operate.

Following the dataset discussion, the chapter explores the programming languages and libraries used to develop and evaluate the models. Python, a widely-used programming language in data science, was employed for its robust support of data manipulation, analysis, and machine learning. The chapter provides an overview of the key libraries used, including those for data handling, visualization, and model implementation. Each library's role and contribution to the project are explained to offer a comprehensive understanding of the technical environment in which the models were developed.

Dataset Description

The dataset used in this study, titled "Soccer players values and their statistics," provides detailed information on football players' market values and performance metrics. This dataset was published in (Stępień, 2021) and is utilized for the analysis in this research.

The market values of footballers are derived from transfermarkt.de, a leading platform known for its football transfer valuations. Transfermarkt.de employs proprietary algorithms to calculate these values, which are influenced by registered users’ votes. Although the exact methodology is not fully transparent, the valuations from transfermarkt.de are highly regarded within the football community and are considered a benchmark by clubs, experts, and journalists. In this study, these market values are assumed to accurately represent the players' market worth.

Additional performance statistics were sourced from fbref.com, which provides a comprehensive array of football statistics accessible to the public. This includes various metrics related to player performance and characteristics.

The dataset covers three football seasons: 2017/18, 2018/19, and 2019/20. This range was chosen because certain variables from fbref.com are not available for earlier seasons, and the 2020/21 season was still ongoing at the time of data collection. The focus is on players from the top five European leagues: the English Premier League, Spanish La Liga, German Bundesliga, Italian Serie A, and French Ligue 1. These leagues were selected due to their high quality and relevance, ensuring that the data is of significant value. Moreover, market value estimates from transfermarkt.de are considered more reliable for these top leagues, which helps mitigate the risk of biased estimations.

Dataset Analysis

The market values of football players for the 2017/2018, 2018/2019 and 2019/20 season, as sourced from transfermarkt.de, reveal significant insights into player valuation distribution. The histogram of market values (Figure 3) demonstrates that the majority of player values are clustered at lower ranges, with a long tail extending towards higher values. This distribution indicates that while many players have market values concentrated around the lower end, a few high-profile players significantly skew the overall distribution.

The top 10 most valuable players for this season are represented as follows: Kylian Mbappé leads with a staggering value of 166.7 million euros, closely followed by Neymar at 162.7 million euros. Lionel Messi, renowned for his exceptional career, is valued at 147.3 million euros. Both Harry Kane and Mohamed Salah are tied with a market value of 140 million euros each. Kevin De Bruyne holds a value of 133.3 million euros, while Sadio Mané is valued at 120 million euros. Raheem Sterling's value stands at 119.3 million euros, and Eden Hazard is valued at 113.3 million euros. Finally, Antoine Griezmann completes the top 10 with a value of 108.7 million euros.

These values illustrate the significant differences in player valuations, highlighting the substantial market worth of top-performing footballers. The distribution of these values is visually represented in the accompanying bar chart, providing a clear depiction of the economic disparity among the elite football players.

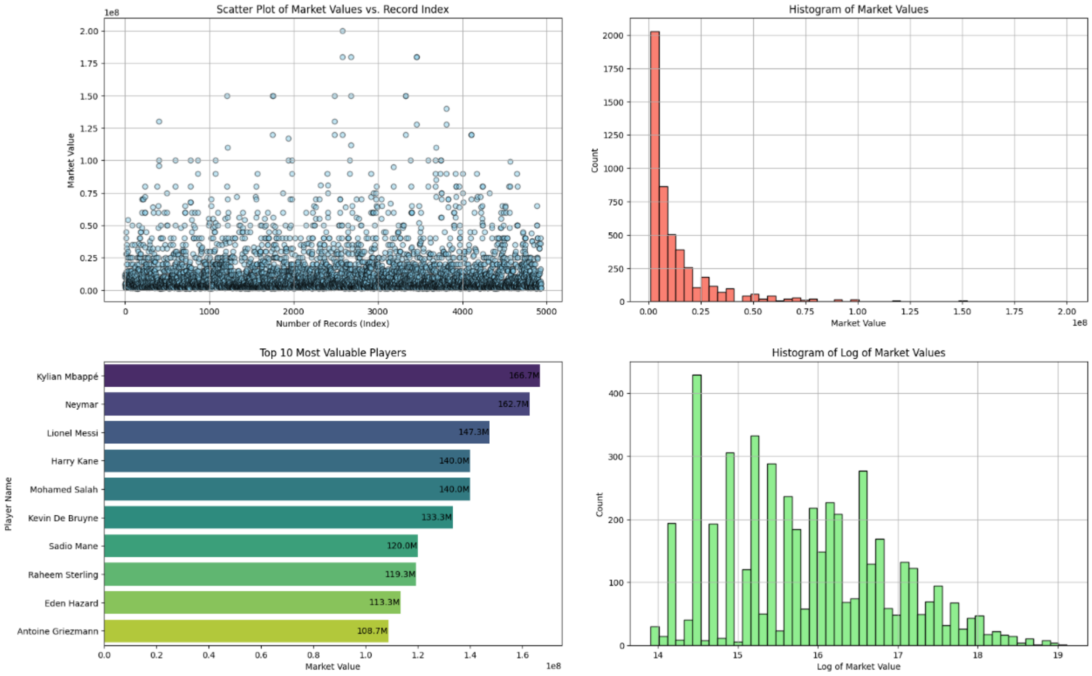


Figure 3: A dashboard displaying the value distribution across all three seasons in the dataset.

Furthermore, the histogram of the market values' logarithm (Figure 3.3) shows a more normalized distribution, illustrating the logarithmic transformation’s effectiveness in managing the skewness present in the raw data. The Jarque-Bera test confirms that the log-transformed values approximate a normal distribution, enhancing the robustness of statistical analysis based on these transformed values.

These plots collectively illustrate the distribution and valuation range of football players, providing a clearer understanding of the market values and their variability within the dataset.

Table 3: Table showing the top 10 most valuable players across all three seasons, along with key attributes.

Player	Age	Position	League
Kylian Mbappé	19	attack - Centre-Forward	Ligue 1
Neymar	26	Forward - Left Winger	Ligue 1
Lionel Messi	31	Forward - Right Winger	La Liga
Harry cane	25	Forward - Centre-Forward	Premier League
Mohammad Salah	26	Forward - Right Winger	Premier League
Kevin De Bruyne	27	midfield - Attacking Midfield	Premier League
Sadio Mang	27	Forward - Left Winger	Premier League
Raheem Sterling	23	attack - Left Winger	Premier League
Eden Hazard	27	Forward - Left Winger	La Liga
Antonie Griezmann	27	attack - Second Striker	La Liga

I begin by examining the ten most valuable players (table 3) and some of their key characteristics. As expected, better on-pitch performance often correlates with higher player value, but several additional factors contribute to a player’s overall value. Among the top 10 most valuable players, Lionel Messi stands out as the only player aged over 30, reflecting his exceptional status in football history. Notably, there are three very young players, each under 21, who may benefit from their potential and future growth, which could influence their current market value.

The analysis also highlights that position plays a significant role in player value. Out of the ten most valuable players, eight are forwards. This aligns with the trend observed in transfermarkt.de, where forwards dominate the list of the most expensive transfers in history. Furthermore, the league in which players compete seems to impact their value, with six of the top 10 players playing in the English Premier League.

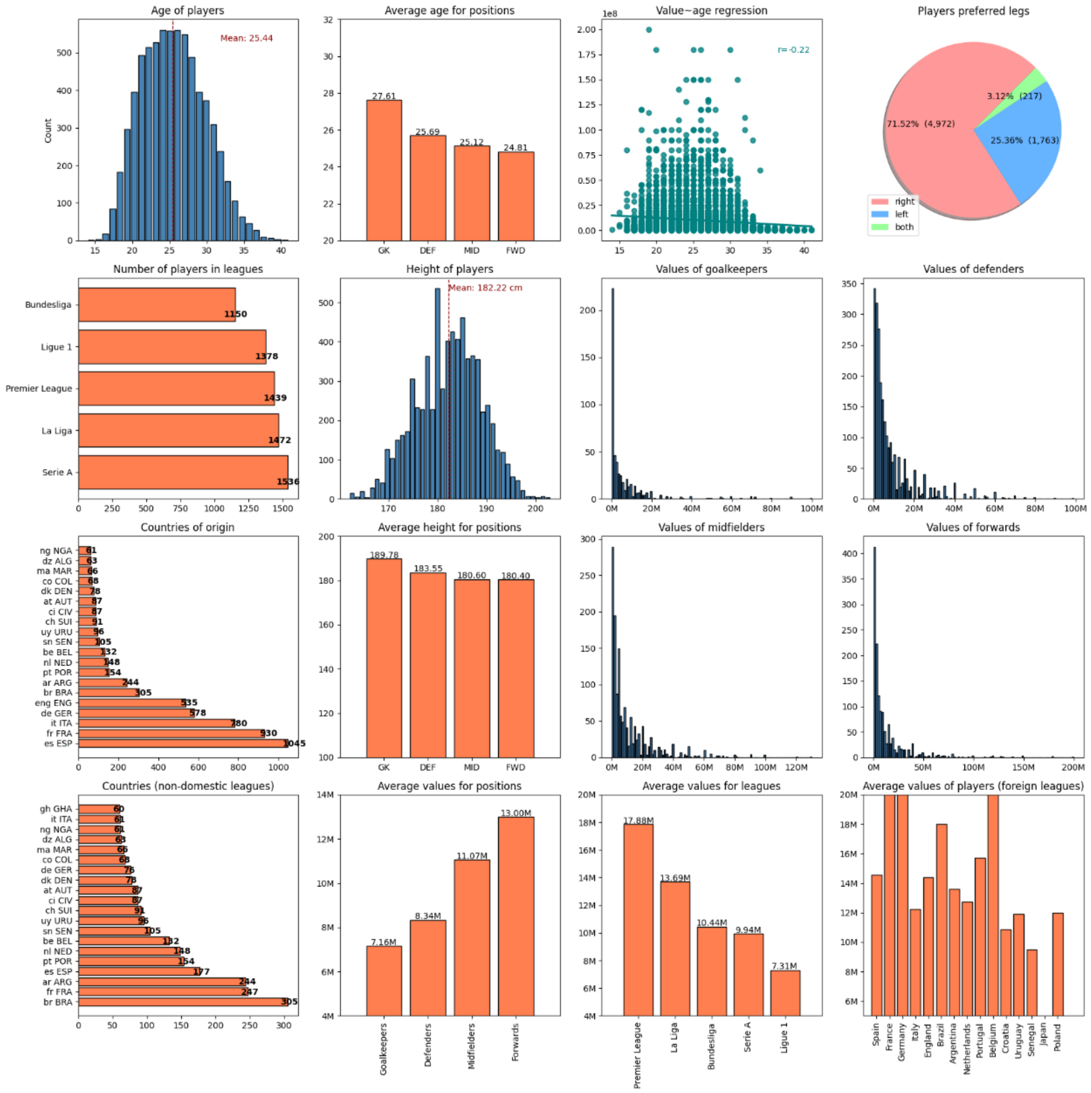


Figure 4: Dashboard illustrating selected characteristics of the dataset and their interrelationships.

Figure 4 illustrates the distribution of key variables in the dataset. The average player age is 25.35 years, with the majority (71.6%) being in their 20s. The youngest player is 14, while the oldest are 41. Goalkeepers are generally older, averaging 27.61 years. The correlation between player value and age is weakly negative (r = 0.23), with a stronger correlation (r ≈ 0.5) for players over 26, indicating a potential non-linear relationship. The Jarque-Bera test shows that the age distribution is not normal, consistent with the known trend of declining market value with age due to diminishing physical attributes. Figure 3.4 also explores player attributes by age using FIFA 21 data, showing general trends in performance metrics.

Despite overall player ratings being consistent across age groups, physical abilities and dribbling skills decline with age, influenced by positional demands. Goalkeepers, who need less mobility, tend to be older on average. The average player height is 182.28 cm, with significant deviations based on position: midfielders and forwards average 180.50 cm, defenders 183.62 cm, and goalkeepers 189.98 cm. Height distribution is skewed, particularly favoring goalkeepers and defenders. Footedness shows that while over 70% of players are right-footed, about 4% are equally adept with both feet, which may affect player value. Nationality data indicates that most players come from countries with top leagues, with Brazilian players leading in foreign leagues. The dataset, with 408 columns including performance metrics and custom variables, necessitates a focused analysis using correlation matrices to identify significant variables and address potential issues such as missing data and outliers.

Dataset Issues

This section addresses concerns related to omitted variables, which could lead to endogeneity bias, reflecting inherent flaws in the dataset. Despite the extensive array of on-pitch statistics, several potentially significant variables are absent. For instance, metrics like average total distance covered per match, which could measure a player's work rate and in-game engagement, are missing. Such statistics, though crucial for football analytics, are not freely available on sources like fbref.com. Another critical variable that comes to mind is a player's injury proneness, which could be inferred from recent injury history. The dataset is also limited to national league statistics, excluding data from international competitions, which might significantly impact the estimations. It is also important to note that the market value of a footballer is inherently endogenous, with several unmeasurable variables, such as a player's innate skills and talent, playing a vital role in the final calculations.

In terms of data preprocessing, I first compiled data from the 2017/18, 2018/19, and 2019/20 seasons. After consolidating and sorting the dataset by player names, I handled missing data and corrected special characters in player names to ensure consistency. I also removed unnecessary columns, such as 'Attendance,' 'birth_year,' and 'MP,' which did not contribute to the analysis. Furthermore, dummy variables were generated for categorical features like 'league,' 'Season,' and 'foot,' and I removed outliers by applying specific criteria: retaining only observations with a market value above one million euros, players with more than five games played, and those with complete data on age and height.

Despite these efforts, the dataset still presents challenges, including several outliers and non-typical observations. For example, there is a notable number of players who played very few games throughout the season or lacked data for variables such as age, height, or preferred foot. Most of these players were deep backups or juniors, though some might be recognized by enthusiasts of specific national leagues. To mitigate the potential negative impact of these variables, I applied the criteria mentioned above, which reduced the dataset from 7,108 to 5,673 observations.

For model training, I plan to create distinct models for the four different positions—goalkeepers, defenders, midfielders, and forwards—by splitting the dataset accordingly. The differences in the number of observations for these positions are expected. Football teams' field 11 players, with only one goalkeeper, typically 3-5 defenders, 2-4 midfielders, and 2-4 forwards. Given the limited number of substitutes allowed per game (three during the used seasons), the relatively small number of goalkeepers in the dataset is unsurprising, and the predominance of defenders aligns with expected team formations.

Programming Language and Libraries

In the development of the predictive models and the overall application, Python was chosen as the primary programming language. Python is widely recognized for its versatility and robust ecosystem, which makes it particularly suited for data analysis and machine learning tasks. Its extensive library support and ease of integration with various tools and frameworks have significantly contributed to the successful implementation of this project.

Python: Python's popularity in the fields of data science and machine learning stems from its readability, flexibility, and the availability of numerous libraries that streamline complex tasks. Python's syntax and structure facilitate clear and efficient coding, which is essential for developing and evaluating predictive models.

Key Libraries Used:

1. **NumPy**:

    - **Purpose**: Provides support for large, multi-dimensional arrays and matrices, along with a collection of mathematical functions to operate on these arrays.

    - **Usage**: NumPy was utilized for numerical operations, including log transformations of data, which are crucial for preparing datasets for modeling.

2. **Pandas**:

   - **Purpose**: Offers data structures and data analysis tools that are ideal for handling and manipulating structured data.

   - **Usage**: Pandas was employed for data preprocessing, including merging datasets, filtering data by position, and managing feature extraction. Its DataFrame structure facilitated efficient data manipulation and analysis.

3. **Statsmodels**:

   - **Purpose**: Provides classes and functions for the estimation of statistical models and performing hypothesis tests.

   - **Usage**: Statsmodels was used for implementing Ordinary Least Squares (OLS) regression models. It enabled the creation of formulas for regression analysis and provided tools for fitting models and generating predictions.

4. **Scikit-learn**:

   - **Purpose**: Offers a range of tools for machine learning, including classification, regression, clustering, and model evaluation.

   - **Usage**: Scikit-learn was integral to implementing machine learning models, such as Decision Trees and Random Forests. It provided functionalities for model training, evaluation, and metrics calculation, including Mean Squared Error and Mean Absolute Error (MAE).

5. **Matplotlib**:

- **Purpose**: Used for creating static, animated, and interactive visualizations in Python.
- **Usage**: Matplotlib was utilized for plotting and visualizing results from model evaluations. It facilitated the creation of graphs to better understand model performance and compare results across different models.

6. **PyQt6**:

- **Purpose**: A set of Python bindings for the Qt application framework, used for developing cross-platform applications with graphical user interfaces.
- **Usage**: Employed to build a user-friendly desktop application for predicting football player market values. PyQt6 was used to design and implement the application's graphical interface, including input forms, navigation, and visualization elements.

By leveraging these libraries, the project effectively addressed the requirements for data handling, model development, and performance evaluation, while PyQt6 ensured a comprehensive and interactive user experience. Each library contributed uniquely to different aspects of the implementation, facilitating a robust approach to building and assessing the predictive models.

In the following chapter, we will discuss the test plan used to evaluate the software, analyze the output from these tests, and assess the project's success and limitations in relation to the aims and objectives outlined in the introduction.

Application Integration and Implementation

The desktop application developed as part of this thesis is designed to predict the market value of football players based on their position and a set of features relevant to each position. The application leverages machine learning models (OLS for goalkeepers, Decision Tree for midfielders, Random Forest for forwards and defenders) to provide accurate predictions. The user interface is built using PyQt6, providing an intuitive and user-friendly experience.

Application Requirements

Functional Requirements

1. **Model Training and Integration:**

    - The application must load pre-trained machine learning models specific to each player position (goalkeeper, midfielder, forward, defender).

    - It must allow the user to select a player position and input relevant features to predict the player's market value.

2. **User Interface:**

    - The application must have a main window with a "Start" button to initiate the process.

    - After starting, the user should be able to select the player's position from a dropdown menu.

    - Upon selecting a position, the application must display a form where the user can input specific player features.

    - The application must include a "Predict Value" button to trigger the prediction process.

    - The predicted market value must be displayed in a user-friendly format (e.g., as a pop-up message).

3. **Prediction Logic:**
   - The application must correctly map the input features to the selected model for prediction.
   - It must convert user inputs into a format compatible with the underlying model for accurate predictions.
   - The prediction results should be converted back from the logarithmic scale to the actual market value in currency format.
4. **Error Handling:**
   - The application must validate user inputs to ensure they are in the correct numeric format.
   - If invalid input is detected, the application should display an error message prompting the user to correct the input.

Non-Functional Requirements

1. **Performance:**
   - The application should load and predict player values within a few seconds to ensure a smooth user experience.
   - It must handle large datasets efficiently when training the models to prevent lag or crashes.
2. **Usability:**
   - The interface should be intuitive, with clear labels and tooltips for each input field to guide the user.

- The application should be accessible to users with basic computer skills and knowledge of football.

3. **Scalability:**

   - The application should be scalable to accommodate additional player positions or features in the future.

4. **Maintainability:**

   - The codebase should be organized and well-documented to allow for easy updates or modifications.
   - The application should be easy to deploy on different systems without requiring extensive setup.

5. **Reliability:**

   - The application should be stable, with minimal risk of crashes or errors during normal operation.
   - The predictions should be reliable and consistent with the performance of the underlying models.

Use Cases

The desktop application developed for predicting football player market values includes a well-defined use case that interacts with both a human actor and the system itself. The primary human actor in this use case is a football scout or analyst who utilizes the application to estimate the market value of players based on their position and individual features. The system, in turn, performs various tasks to support this functionality, including preparing data, training models, and visualizing results.

Use Case Diagram Description

According to the figure below (Figure 5), The use case diagram for the application features two primary actors:

1. **Football Scout or Analyst (Human Actor)**
2. **System (System Actor)**

Football Scout or Analyst is connected to the following use cases:

- **Launch Application**: The scout or analyst initiates the application.
- **Select Player Position**: The scout selects the specific position (e.g., goalkeeper, defender, midfielder, forward) of the player whose market value they wish to predict.
- **Enter Player Features**: The scout inputs relevant player features, such as age, league, performance statistics, etc.
- **Predict Market Value**: The scout triggers the prediction process, receiving the estimated market value of the player.

System is connected to the following use cases:

- **Prepare Data for Prediction Process**: The system preprocesses the input data, ensuring it is formatted and ready for the prediction model.
- **Training Predictive Models**: The system trains the necessary predictive models based on the input data and selected player position.
- **Visualize Prediction Result**: The system generates a visual representation of the predicted market value, which is displayed to the scout.

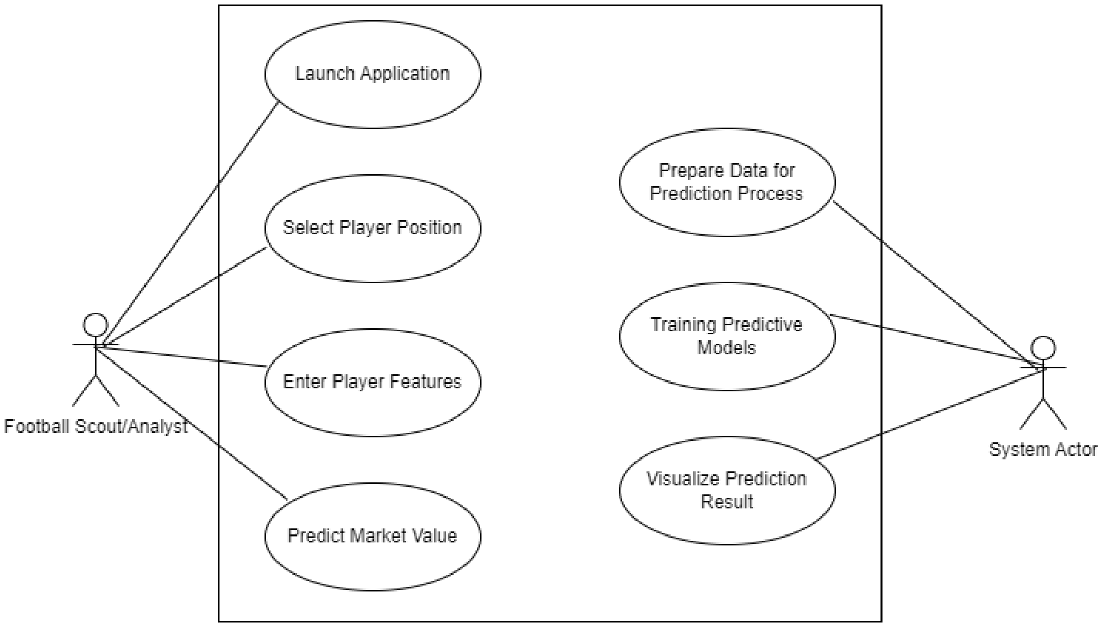


Figure 5: Use Case Diagram for Football Player Market Value Prediction Application.

Scenarios

Scenario 1: Predicting Market Value for a Goalkeeper

- **Actor**: Football Scout or Analyst
- **Steps**:
  1. The scout launches the application.
  2. The scout selects "Goalkeeper" from the position selection screen.
  3. The scout enters relevant features, such as the goalkeeper’s age, league, number of clean sheets, etc.
  4. The scout clicks the "Predict" button.
  5. The system prepares the data, applies the trained Ordinary Least Squares (OLS) model, and predicts the goalkeeper’s market value.
  6. The system visualizes the prediction result, displaying the estimated value on the screen.

Scenario 2: Predicting Market Value for a Midfielder

- **Actor**: Football Scout or Analyst
- **Steps**:
    1. The scout launches the application.
    2. The scout selects "Midfielder" from the position selection screen.
    3. The scout enters relevant features, such as the midfielder’s passing accuracy, goals scored, assists, etc.
    4. The scout clicks the "Predict" button.
    5. The system prepares the data, applies the trained Decision Tree model, and predicts the midfielder’s market value.
    6. The system visualizes the prediction result, displaying the estimated value on the screen.

Scenario 3: Training Predictive Models

- **Actor**: System
- **Steps**:
    1. The system receives the player features entered by the scout.
    2. The system identifies the selected player position.
    3. The system retrieves the appropriate predictive model (e.g., Random Forest for defenders).
    4. The system trains the model using the appropriate data.
    5. The system stores the trained model for use in predicting market values.
    6. The system uses the trained model for prediction and visualizes it.

These scenarios exemplify how the application serves as a practical tool for scouts and analysts, enabling them to predict football player market values efficiently through a user-friendly interface.

Application Design and User Experience

The application interface is designed to be intuitive and user-friendly, ensuring that users can easily navigate through the prediction process. Upon launching the application, users are greeted with a start page (Figure 6) that features a clean, minimalist design, inviting them to begin the prediction process. The start button on this page, prominently displayed and easily accessible, leads users to the position selection screen (Figure 7). This screen allows users to choose the player's position they wish to evaluate, with options for Goalkeeper, Defender, Midfielder, and Forward.

For each selected position, a tailored input form is presented (Figures 8-11), where users can enter specific features relevant to that position. These features were determined during the model training phase, ensuring that the input forms are closely aligned with the data requirements of the underlying predictive models. To enhance usability, each form includes descriptions of the features, guiding the user in providing the necessary data. The predict button, prominently displayed on each form, triggers the prediction process, utilizing the best-performing model for the respective position to calculate the player's market value.

The application integrates different machine learning models, each chosen based on its superior performance for a specific player position:

- **Goalkeepers**: The application uses the Ordinary Least Squares (OLS) model, which performed best for predicting the values of goalkeepers.

- **Defenders**: The Random Forest model, which excelled in this category, is used to predict the values for defenders.
- **Midfielders**: The Decision Tree model was chosen as the best performer and is employed to predict midfielders' values.
- **Forwards**: The Random Forest model, which demonstrated superior accuracy, is used for predicting forward players' values.

After entering the necessary features, users can view the predicted market value, which is displayed in a pop-up window. This value is calculated using the logarithm of market values, ensuring the prediction is both accurate and reliable.

Visual Design and Aesthetic Choices

The visual elements of the application were carefully selected to create an engaging and user-friendly experience. The overall design philosophy emphasizes simplicity and clarity, ensuring that users can navigate through the application with ease while enjoying a visually appealing interface.

- **Background Images**: High-quality background images are used throughout the application to create an immersive environment. These images are relevant to the football context, subtly reinforcing the application's theme without distracting the user. The start page background sets the tone, providing a professional and polished look that engages users from the outset.
- **Color Scheme and Gradients**: The application uses a consistent color scheme with deep blues and neutral tones. These colors evoke a sense of trust and reliability, which are essential in a predictive analytics context. The dark blue gradients in the background contribute to a modern aesthetic while guiding the user’s focus toward the central content. The color scheme

also ensures that text and input fields are easily readable, enhancing the overall usability.

- **Typography and Layout**: Modern, clean fonts are used throughout the application, supporting the minimalist design approach. The layout is carefully designed to ensure that all interactive elements, such as buttons and input fields, are easily accessible. The "Predict Value" button on each form is prominently displayed, with a design that encourages user interaction. The spacing and padding around elements are adjusted to maintain an organized and professional interface.
- **Interactive Elements**: Buttons and input fields are designed not only for functionality but also for aesthetics. Rounded corners, subtle shadows, and smooth animations make these elements stand out without overwhelming the user. The responsive hover effects and click animations provide immediate feedback, enhancing the overall user experience.

These visual elements were chosen with the primary goal of enhancing the user experience. By creating a visually cohesive and aesthetically pleasing interface, the application ensures that users can focus on entering data and generating predictions without unnecessary distractions. The careful balance of functionality and design contributes to an experience that is both efficient and enjoyable.

The figures below provide a visual representation of the application's pages:

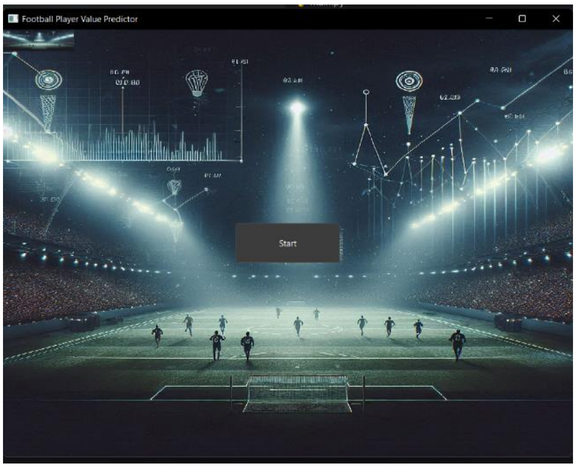


Figure 6: The start page of the application.

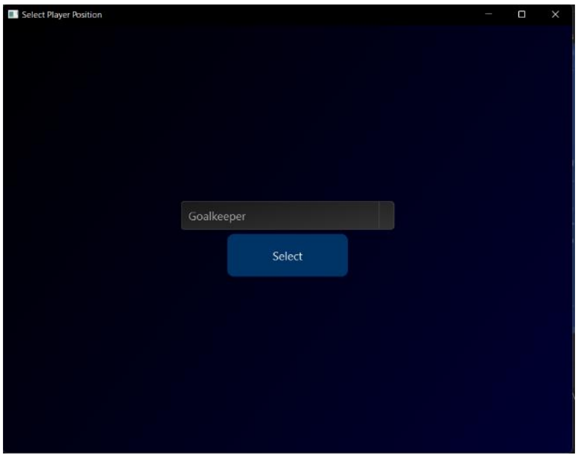
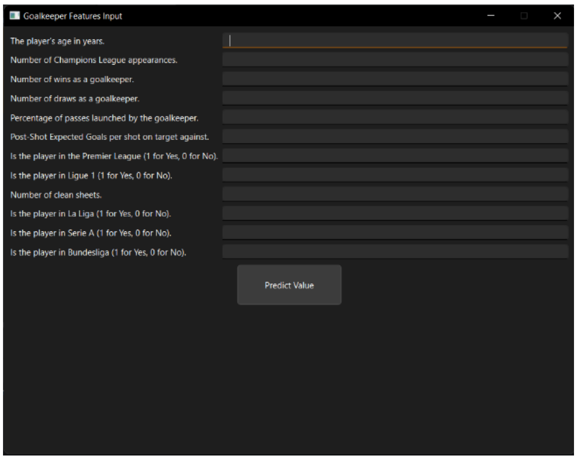


Figure 7: Position selection screen.



The player's age in years.

Number of Champions League appearances.

Number of wins as a goalkeeper.

Number of draws as a goalkeeper.

Percentage of passes launched by the goalkeeper.

Post-Shot Expected Goals per shot on target against.

Is the player in the Premier League (1 for Yes, 0 for No).

Is the player in Ligue 1 (1 for Yes, 0 for No).

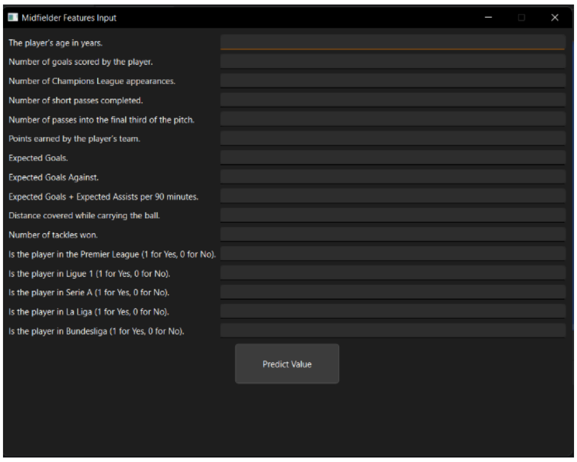
Number of clean sheets.

Is the player in La Liga (1 for Yes, 0 for No).

Is the player in Serie A (1 for Yes, 0 for No).

Is the player in Bundesliga (1 for Yes, 0 for No).

Figure 8:Input form for Goalkeepers.



The player's age in years.

Number of goals scored by the player.

Number of Champions League appearances.

Number of short passes completed.

Number of passes into the final third of the pitch.

Points earned by the player's team.

Expected Goals.

Expected Goals Against.

Expected Goals + Expected Assists per 90 minutes.

Distance covered while carrying the ball.

Number of tackles won.

Is the player in the Premier League (1 for Yes, 0 for No).

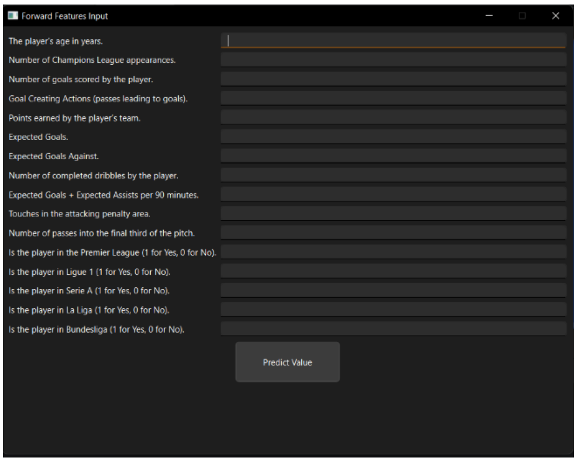
Is the player in Ligue 1 (1 for Yes, 0 for No).

Is the player in Serie A (1 for Yes, 0 for No).

Is the player in La Liga (1 for Yes, 0 for No).

Is the player in Bundesliga (1 for Yes, 0 for No).

Figure 9: Input form for Midfielders.



The player's age in years.

Number of Champions League appearances.

Number of goals scored by the player.

Goal Creating Actions (passes leading to goals).

Points earned by the player's team.

Expected Goals.

Expected Goals Against.

Number of completed dribbles by the player.

Expected Goals + Expected Assists per 90 minutes.

Touches in the attacking penalty area.

Number of passes into the final third of the pitch.

Is the player in the Premier League (1 for Yes, 0 for No).

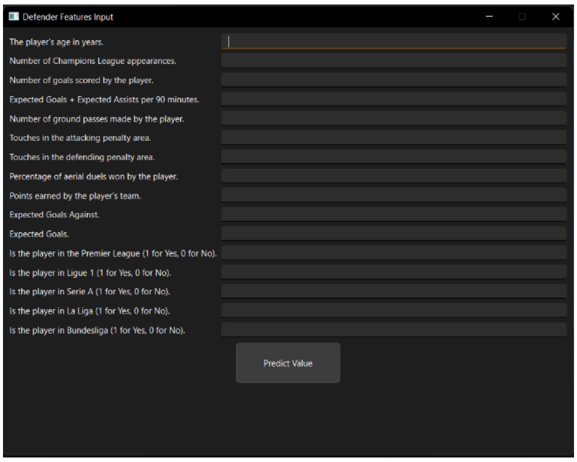
Is the player in Ligue 1 (1 for Yes, 0 for No).

Is the player in Serie A (1 for Yes, 0 for No).

Is the player in La Liga (1 for Yes, 0 for No).

Is the player in Bundesliga (1 for Yes, 0 for No).

Figure 10: Input form for Forwards.



The player's age in years.

Number of Champions League appearances.

Number of goals scored by the player.

Expected Goals + Expected Assists per 90 minutes.

Number of ground passes made by the player.

Touches in the attacking penalty area.

Touches in the defending penalty area.

Percentage of aerial duels won by the player.

Points earned by the player's team.

Expected Goals Against.

Expected Goals.

Is the player in the Premier League (1 for Yes, 0 for No).

Is the player in Ligue 1 (1 for Yes, 0 for No).

Is the player in Serie A (1 for Yes, 0 for No).

Is the player in La Liga (1 for Yes, 0 for No).

Is the player in Bundesliga (1 for Yes, 0 for No).

Figure 11: Input form for Defenders.

By combining the predictive power of the best models with a well-designed user interface, this application serves as a practical tool for estimating football player market values, making it useful for analysts, scouts, and football enthusiasts alike.

CHAPTER 5

EVALUATION

Introduction

This chapter details the process of training and evaluating predictive models for estimating football player market values across different positions. We utilized three regression models—Ordinary Least Squares, Decision Tree Regressor, and Random Forest Regressor—to assess their effectiveness in predicting player values for goalkeepers, defenders, midfielders, and forwards.

We describe the preparation of datasets, including the application of log transformations to address data skewness, and evaluate model performance using Mean Absolute Error and Mean Squared Error. The chapter includes comparisons of model performance across positions, supported by tables and visualizations.

Additionally, we discuss the integration of the best-performing models into a user-friendly desktop application developed with Python and PyQt6. This application allows users to predict player market values based on specific features, demonstrating the practical application of our models.

Evaluation Metrics

To assess the performance of the predictive models, evaluation metrics are used. These metrics quantify the accuracy and reliability of the predictions, providing a basis for comparison between different models. The following metrics are used in this study:

- **Mean Squared Error (MSE):** The average of the squared differences between the observed and predicted values. MSE penalizes larger errors more severely, making it sensitive to outliers (Makridakis, Wheelwright and Hyndman, 2008). It is calculated as:

$$MSE = \frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

- **Mean Absolute Error (MAE):** The average of the absolute differences between the observed and predicted values. MAE provides a straightforward measure of prediction accuracy and is less sensitive to outliers compared to MSE (Willmott and Matsuura, 2005; Chai and Draxler, 2014). It is calculated as:

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i|$$

These metrics are crucial for evaluating the effectiveness of the models in predicting the market value of football players and for comparing the performance of different regression techniques.

Model Training and Evaluation for Positions

In the initial step of our analysis, we aimed to identify the best-performing model for predicting player performance in different football positions. We employed three different models—Ordinary Least Squares, Decision Tree Regressor, and Random Forest Regressor—and evaluated their effectiveness using specific datasets for each position. The following sections describe the process in detail.

For each model, we began by preparing the dataset for training and evaluation. The dataset used for each position included various features relevant to the position. Importantly, we added a log-transformed target variable (ln_value) to address potential skewness in the distribution of the target variable value. This transformation helps stabilize variance and make the model's assumptions more realistic.

We defined three functions to handle the training and evaluation of each model. Each function follows a similar process: splitting the data into training and test sets, training the model on the training set, making predictions on the test set, and then calculating evaluation metrics. Below is a description of each function:

The OLS model is trained using a formula-based approach. It provides an interpretative result but may struggle with non-linear relationships and interactions. The Decision Tree model captures non-linear relationships and interactions between features but can be prone to overfitting. The Random Forest model, an ensemble method, combines multiple decision trees to improve accuracy and reduce overfitting. This diverse set of features helps in capturing various aspects of a goalkeeper's performance.

Model Evaluation

To evaluate the performance of each predictive model—Ordinary Least Squares, Decision Tree Regressor, and Random Forest Regressor—we calculated the Mean Absolute Error and Mean Squared Error separately for each player position: goalkeeper, defender, midfielder, and forward. These metrics provide insight into how well each model predicts player market values across different positions.

For each position, we trained the models and then assessed their predictions on the test set using MAE and MSE. The results revealed varying levels of performance across the positions, with the Random Forest Regressor generally achieving the lowest MAE and MSE, indicating its superior accuracy and robustness.

To visualize the results, we plotted the MAE and MSE for each model across all player positions. These plots allow us to compare the prediction errors for each model, helping to identify the best-performing model for each specific position. By examining these plots, we can see which model consistently produces predictions that are closest to the actual values, particularly highlighting the model that performs best overall.

After selecting the best model for each position based on the lowest errors, we also presented a sample of predicted versus actual values for that model within each position. This demonstration further illustrates the practical prediction capabilities of the chosen models, providing clear evidence of their effectiveness in predicting player market values.

Goalkeepers

Model Comparison

Table 4: MAE and MSE of Predictive Models for Goalkeepers.

Model	**MAE**	**MSE**
OLS	0.4467	0.2769
Decision tree	0.6860	0.7770
Random forest	0.5668	0.4407

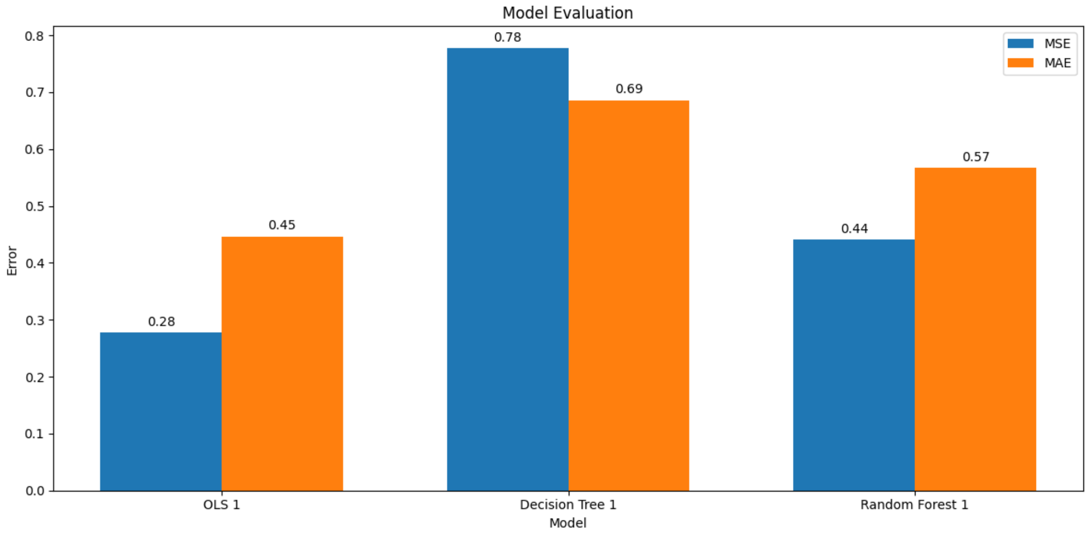


Figure 12: MAE and MSE Comparison for Goalkeeper Models.

Best Model

Error analysis and plot review reveal that the OLS model is the top performer for goalkeepers. To further evaluate accuracy, we compared the actual versus predicted logarithmic values for each model. The OLS model, which emerged as the best, is highlighted with sample results below, presented in logarithmic form (ln).

Table 5: sample of predicted values vs actual values for best model for Goalkeepers.

Actual	Predicted
14.5086	14.9220
15.201805	15.5259
16.4545	15.6656
14.9141	15.0136

14.9143	15.3770
15.8949	15.8054
14.3751	14.9758
14.9141	15.4502
15.4249	15.7385
14.9141	15.3721

Defenders

Model Comparison

Table 6: MAE and MSE of Predictive Models for Defenders.

Model	**MAE**	**MSE**
OLS	0.3376	0.4560
Decision tree	0.6564	0.6016
Random forest	0.2957	0.4180

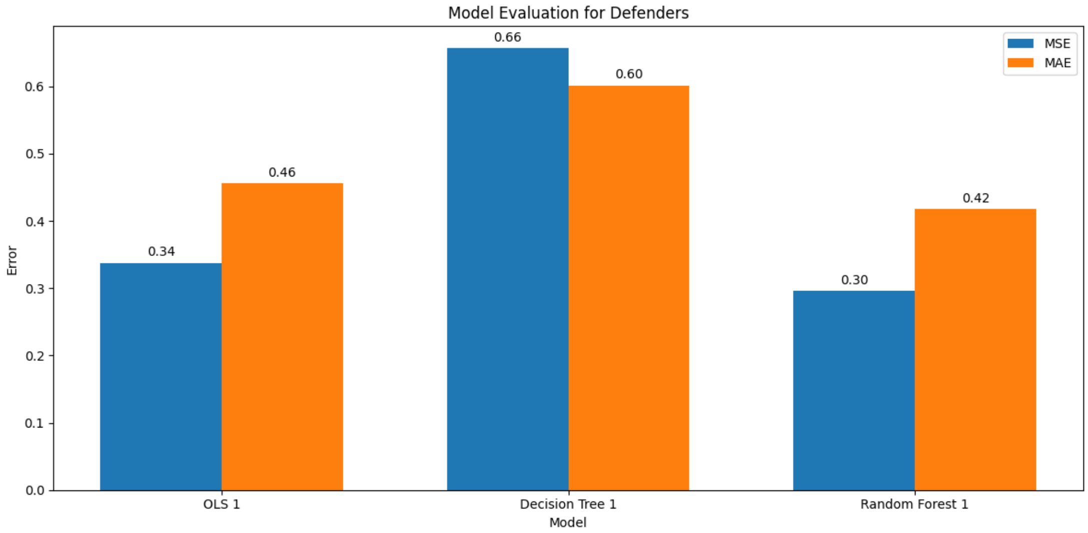


Figure 13: MAE and MSE Comparison for Defender Models.

Best Model

By evaluating the errors and interpreting the plot, it becomes clear that the Random Forest model performs optimally for defenders. We also reviewed the actual versus predicted logarithmic values for a visual performance assessment of each model. The Random Forest, identified as the best model, is illustrated with sample results below, where values are in logarithmic form (ln).

Table 7: sample of predicted values vs actual values for best model for Defenders.

Actual	Predicted
16.648724	15.763265
15.201805	15.580780
15.424948	15.004404
15.830414	15.495847

16.118096	15.357446
14.914123	15.095743
17.909855	17.781847
15.424948	15.264583
15.424948	16.074725
15.201805	15.376045

Midfielders

Model Comparison

Table 8: MAE and MSE of Predictive Models for Midfielders.

Model	**MAE**	**MSE**
OLS	0.3442	0.4575
Decision tree	0.0783	0.1004
Random forest	0.0992	0.2290

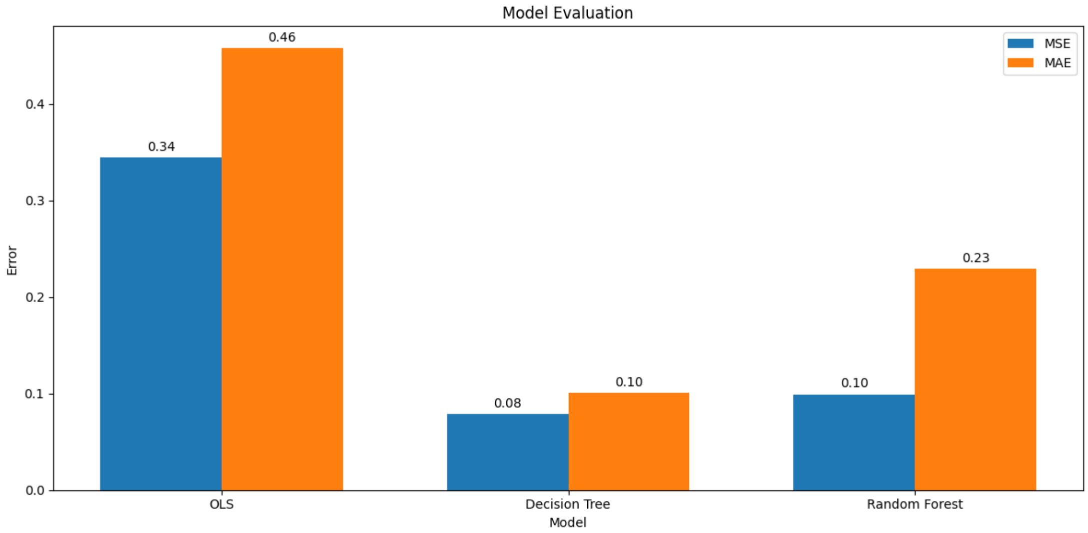


Figure 14: MAE and MSE Comparison for Midfielder Models.

Best Model

An analysis of the errors and the plot indicates that the Decision model excels for midfielders. We compared the actual and predicted logarithmic values across models to assess their accuracy visually. The Decision Tree, emerging as the best, is showcased with sample results below. The values are presented in logarithmic form (ln).

Table 9: sample of predicted values vs actual values for best model for Midfielders.

Actual	Predicted
15.201805	15.201805
15.761421	15.761421
16.523561	16.118096
15.830414	15.495847

16.705882	16.705882
15.830414	15.830414
15.894952	15.894952
18.315320	18.315320
14.914123	14.914123
14.077875	14.077875

Forwards

Model Comparison

Table 10: MAE and MSE of Predictive Models for Forwards.

Model	**MAE**	**MSE**
OLS	0.2760	0.4243
Decision tree	0.8090	0.6915
Random forest	0.3697	0.4814

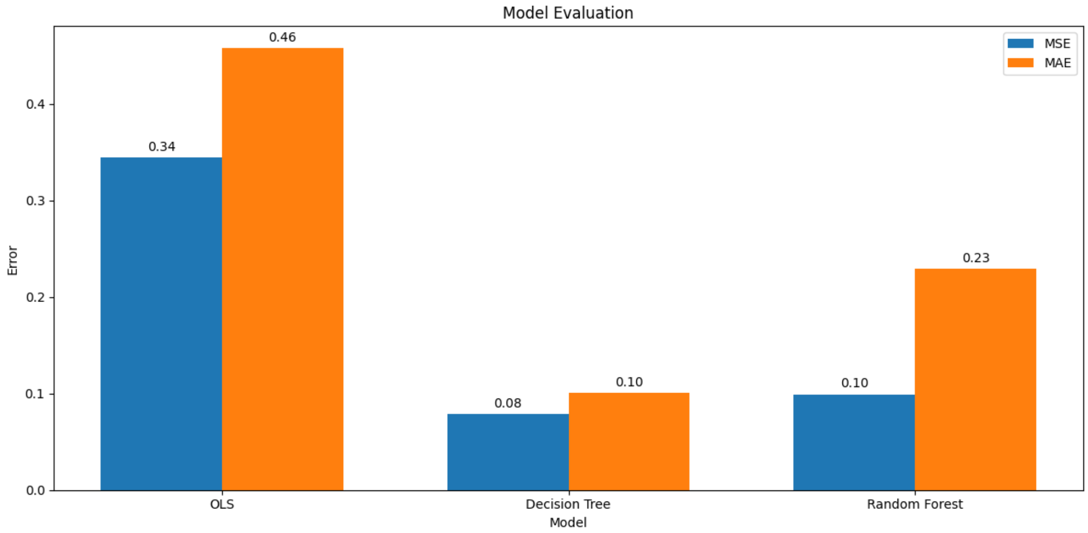


Figure 15: MAE and MSE Comparison for Forward Models.

Best Model

Upon reviewing the errors and analyzing the plot, the Random Forest model demonstrates superior performance for forwards. We also examined the actual versus predicted logarithmic values for each model to visually assess accuracy. Below are sample results for the Random Forest, which proved to be the most effective. Note that the values are displayed in logarithmic form (ln).

Table 11: sample of predicted values vs actual values for best model for Forwards.

Actual	Predicted
15.424948	15.645521
18.132999	17.398586
14.508658	15.362156
16.300417	15.820255
16.811243	15.980076

18.826146	18.633480
16.118096	15.901093
15.201805	15.970527
16.012735	15.415285
16.811243	16.727992

CHAPTER 6

CONCLUSIONS / FUTURE WORK

Introduction

This chapter provides a comprehensive summary of the work undertaken in this project, reflecting on the research questions and objectives that guided the study. It evaluates the success of the project in meeting its goals and acknowledges the limitations encountered during the development process. In addition, this chapter offers recommendations for future work that could enhance the application and its predictive capabilities. The discussion also extends to the professional, social, ethical, and legal implications of the project, considering the broader impact and responsibilities associated with deploying machine learning models in a real-world context.

Summary and Analysis

This research aimed to develop a desktop application capable of predicting the market values of football players based on their performance metrics, utilizing machine learning models tailored to specific player positions. By addressing the initial research questions and hypotheses, this project has successfully demonstrated that different machine learning models can be effectively employed to predict player market values with a reasonable degree of accuracy.

The models developed—Ordinary Least Squares, Decision Tree Regressor, and Random Forest Regressor—were each evaluated across four player positions: goalkeepers, defenders, midfielders, and forwards. The results showed that the Random Forest Regressor performed exceptionally well for defenders and forwards, while the Decision Tree Regressor was more suitable for midfielders. The OLS model proved most effective for goalkeepers. These findings support the hypothesis that model performance varies significantly depending on the specific attributes and performance metrics associated with different player positions.

The application was successfully developed and integrated with these models, providing users with a practical tool to predict player market values. The project's objectives—to create a user-friendly interface, ensure accurate predictions, and offer valuable insights into player valuations—were largely achieved.

Project Success and Limitations

The project achieved its primary objectives by delivering a functional desktop application capable of predicting football player market values. The accuracy of the predictions, particularly for certain positions, indicates the effectiveness of the chosen models and the preprocessing techniques applied. The application's design and user interface were also successfully implemented, resulting in an intuitive and aesthetically pleasing user experience.

However, the project faced several limitations. The accuracy of the predictions, while reasonable, could be improved, particularly for certain player positions. The models were trained on a limited dataset, which may have restricted their ability to generalize to new data. Additionally, the omission of certain potentially influential variables due to data availability or preprocessing decisions could have impacted the models' performance.

Furthermore, the application currently predicts market values based solely on historical data and performance metrics, without considering external factors such as market trends, player injuries, or transfer rumors, which can significantly influence market values. The exclusion of these variables limits the model's ability to capture the full complexity of player valuation.

Recommendations for Future Work

Building on the accomplishments of this project, several recommendations for future work are proposed:

1. Expand the Dataset: Future research could benefit from a more extensive dataset, including data from additional seasons and leagues. This would improve the models' ability to generalize and potentially enhance prediction accuracy.

2. Incorporate External Factors: To improve the predictive power of the models, it would be valuable to include external factors such as market trends, economic conditions, player injuries, and transfer rumors. These variables could provide a more comprehensive view of player market values.

3. Model Optimization: Further tuning and optimization of the models could be conducted to improve their performance. Exploring more advanced machine learning techniques, such as ensemble methods or deep learning models, could also be beneficial.

4. Real-Time Prediction: Implementing real-time data integration, where the application could update predictions based on the latest available data, would enhance its practical utility.

5. Cross-Platform Development: Expanding the application to support multiple platforms, including mobile devices, could increase its accessibility and user base.

In conclusion, while the project has successfully achieved its primary goals, there are opportunities for further improvement and expansion. By addressing the limitations and recommendations outlined, future work can build on this foundation to create even more accurate and comprehensive tools for football player market value prediction. As the application evolves, careful consideration of the associated professional, social, ethical, and legal issues will remain essential.

REFERENCES

Abramson, N., Braverman, D. and Sebestyen, G. (1963) ‘Pattern recognition and machine learning’, *IEEE Transactions on Information Theory*, 9(4), pp. 257–261.

Behravan, I. and Razavi, S.M. (2021) ‘A novel machine learning method for estimating football players’ value in the transfer market’, *Soft Computing*, 25(3), pp. 2499–2511. Available at: https://doi.org/10.1007/s00500-020-05319-3.

Bialkowski, A. *et al.* (2014) ‘Win at home and draw away: Automatic formation analysis highlighting the differences in home and away team behaviors’, in *Proceedings of 8th annual MIT sloan sports analytics conference*. Citeseer, pp. 1–7. Available at: https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=c932441f8ecc9348e2dd54ec749984c8f4abf224 (Accessed: 26 August 2024).

Breiman, L. (1996) ‘Bagging predictors’, *Machine Learning*, 24(2), pp. 123–140. Available at: https://doi.org/10.1007/BF00058655.

Breiman, L. (2017) *Classification and regression trees*. Routledge. Available at: https://www.taylorfrancis.com/books/mono/10.1201/9781315139470/classification-regression-trees-leo-breiman (Accessed: 14 August 2024).

Chai, T. and Draxler, R.R. (2014) ‘Root mean square error (RMSE) or mean absolute error (MAE)?–Arguments against avoiding RMSE in the literature’, *Geoscientific model development*, 7(3), pp. 1247–1250.

Decroos, T. *et al.* (2020) ‘VAEP: An objective approach to valuing on-the-ball actions in soccer’, in *IJCAI*, pp. 4696–4700. Available at: https://www.ijcai.org/proceedings/2020/0648.pdf (Accessed: 26 August 2024).

Elahi, M., Pandey, S. and Malhi, S.S. (2023) ‘Market Value Prediction of Football Players’. Rochester, NY. Available at: https://doi.org/10.2139/ssrn.4485449.

Gudmundsson, J. and Horton, M. (2018) ‘Spatio-Temporal Analysis of Team Sports’, *ACM Computing Surveys*, 50(2), pp. 1–34. Available at: https://doi.org/10.1145/3054132.

Hastie, T., Tibshirani, R. and Friedman, J. (2001) ‘The elements of statistical learning. Springer series in statistics’, *New York, NY, USA* [Preprint].

Huang, C. and Zhang, S. (2023) ‘Explainable artificial intelligence model for identifying Market Value in Professional Soccer Players’, *arXiv e-prints*, p. arXiv-2311.

James, G. *et al.* (2023) *An Introduction to Statistical Learning: with Applications in Python*. Cham: Springer International Publishing (Springer Texts in Statistics). Available at: https://doi.org/10.1007/978-3-031-38747-0.

Lasek, J., Szlávik, Z. and Bhulai, S. (2013) ‘The predictive power of ranking systems in association football’, *International Journal of Applied Pattern Recognition*, 1(1), p. 27. Available at: https://doi.org/10.1504/IJAPR.2013.052339.

Makridakis, S., Wheelwright, S.C. and Hyndman, R.J. (2008) *Forecasting methods and applications*. John wiley & sons.

Mitchell, T. (1997) ‘Introduction to machine learning’, *Machine learning*, 7, pp. 2–5.

Murphy, K.P. (2012) *Machine Learning–A probabilistic Perspective*. The MIT Press.

Pappalardo, L. *et al.* (2019) ‘PlayeRank: Data-driven Performance Evaluation and Player Ranking in Soccer via a Machine Learning Approach’, *ACM Transactions on Intelligent Systems and Technology*, 10(5), pp. 1–27. Available at: https://doi.org/10.1145/3343172.

Pappalardo, L. and Cintia, P. (2018) ‘QUANTIFYING THE RELATION BETWEEN PERFORMANCE AND SUCCESS IN SOCCER’, *Advances in Complex Systems*, 21(03n04), p. 1750014. Available at: https://doi.org/10.1142/S021952591750014X.

Royce, W.W. (1970) ‘Managing the development of large software systems. proceedings of IEEE WESCON’, *Los Angeles*, pp. 328–388.

Schölkopf, B., Platt, J. and Hofmann, T. (2007) ‘TrueSkill^{TM}: A Bayesian Skill Rating System’. Available at: https://ieeexplore.ieee.org/abstract/document/6287323/ (Accessed: 20 July 2024).

Schulte, O. *et al.* (2017) ‘Apples-to-apples: Clustering and ranking NHL players using location information and scoring impact’, in *Proceedings of the MIT Sloan Sports Analytics Conference*.

Shearer, C. (2000) ‘The CRISP-DM model: the new blueprint for data mining’, *Journal of data warehousing*, 5(4), pp. 13–22.

Stępień, R. (2021) *Modelling football players values on the transfer market and their determinants using robust regression models*. Warsaw School of Economics. Available at: https://github.com/RSKriegs/Modelling-Football-Players-Values-on-a-Transfer-Market/blob/main/RS82640%20Modelling%20Footballers%20Values%20on%20a%20Transfer%20Market%20.pdf (Accessed: 5 July 2024).

The 100 most expensive football transfers of all time | Goal.com UK (2023). Available at: https://www.goal.com/en-gb/news/100-most-expensive-football-transfers-all-time/ikr3oojohla51fh9adq3qkwpu (Accessed: 19 August 2024).

Willmott, C.J. and Matsuura, K. (2005) ‘Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance’, *Climate research*, 30(1), pp. 79–82.

Yiğit, A.T., Samak, B. and Kaya, T. (2020) ‘Football player value assessment using machine learning techniques’, in *Intelligent and Fuzzy Techniques in Big Data Analytics and Decision Making: Proceedings of the INFUS 2019 Conference, Istanbul, Turkey, July 23-25, 2019*. Springer, pp. 289–297. Available at: https://link.springer.com/chapter/10.1007/978-3-030-23756-1_36 (Accessed: 20 July 2024).