Exploring Football Player Salary Prediction Using Random Forest: Leveraging Player Demographics and Team Associations

Riyadh Abdulhadi M Aljohani[1,*], Abdulaziz Amir Alnahdi[2]

[1,2]*Information Science Department, King Abdulaziz University, Jeddah, Saudi Arabia*

(Received: June 10, 2025; Revised: July 20, 2025; Accepted: October 31, 2025; Available online: December 10, 2025)

Abstract

This paper explores the prediction of football player salaries using a Random Forest (RF) Regressor model, leveraging player demographics and team associations as key features. The dataset consists of 684 football players, including variables such as age, nationality, position, team, weekly salary, and annual salary. The study applies Exploratory Data Analysis (EDA) to understand the distribution of these features and identify patterns within the dataset. Data preprocessing involves handling missing values, one-hot encoding categorical variables, and splitting the dataset into training and testing sets. The RF model is trained on the preprocessed data, and its performance is evaluated using common regression metrics, including R-squared (R^2), Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE). The results show that the model explains approximately 48.5% of the variance in player salaries, with an MAE of £1.92 million and an RMSE of £2.82 million. Key predictors of salary include player age, position, nationality, and team. The analysis of feature importance reveals that categorical variables such as Nation and Team have a significant impact on salary predictions. However, the model's performance is constrained by the lack of more granular data, such as player performance metrics or external economic factors. This research provides valuable insights for football team management, helping teams understand which factors contribute to salary setting and enabling more informed decisions in player recruitment and contract negotiations. It also highlights the potential for sponsorships to target players based on these predictive attributes. Future work could explore the integration of more advanced machine learning techniques and additional player data to improve predictive accuracy and model robustness.

Keywords: Football Player Salary Prediction, Random Forest, Machine Learning, Exploratory Data Analysis, Feature Importance

1. Introduction

The prediction of football player salaries is an increasingly relevant area of economic and sports analytics research. With the financial stakes involved in player contracts, accurate salary prediction models provide valuable insights for clubs, players, and agents. This complexity arises from various interdependent factors that influence player wages, including performance statistics, market dynamics, and social behaviors within the team environment. It is vital to consider how quantitative methods not only furnish data-driven insights but may also impact behavioral dynamics within a team, ultimately influencing player performance [1]. The application of machine learning techniques to enhance salary prediction accuracy has gained prominence in contemporary research. Algorithms such as regression analysis and classification models have been leveraged to predict future salaries for active players based on historical data and multiple influencing variables [2], [3]. Scholars like Matbouli and Alghamdi [3] emphasize the holistic approach required for salary prediction, which should incorporate economic and occupational characteristics. They highlight how utilizing and refining existing models enables a deeper understanding of how different factors interact [3]. In turn, incorporating such algorithms significantly reduces overfitting risks and improves the models' predictive power, as observed in recent comparisons of deep-learning algorithms with conventional methods [4].

Furthermore, the influence of club revenues on player salaries cannot be overlooked, as it is paramount to the economic landscape of leagues such as the English Premier League (EPL). Research indicates that television revenues represent the most substantial source of income for club operations, closely linked to player salaries and transfer expenses [5]. The COVID-19 pandemic provided a unique case study, demonstrating how a downturn in market conditions could

[*]**Corresponding author: Riyadh Abdulhadi M Aljohani (raljohani0265@stu.kau.edu.sa)**
DOI: https://doi.org/10.47738/ijaim.v5i4.115

affect these revenue streams and, by extension, future salary predictions. The interconnectedness of club revenues and player salaries illustrates a critical factor in developing accurate salary prediction models. In terms of practical applications, one study leveraged data from FIFA video games in conjunction with machine learning techniques to formulate a robust salary prediction model. This model offers an objective quantifiable approach to estimating player market values, which could be instrumental during negotiations between clubs and agents [6]. The insights gleaned from this type of analysis showcase the potential to streamline contract negotiations and enhance transparency regarding player worth in the transfer market.

The need for accurate salary prediction models in football is underscored by several critical factors within the economic, social, and performance dynamics of the sport. Traditional salary structures are often based on various subjective measures, leading to inconsistencies and disparities within team compositions, which can affect overall performance and morale. It has been documented that salary inequality can have detrimental impacts on player performance, with the negative effects becoming more pronounced as salary differences increase [1]. This highlights the necessity of establishing models that not only provide reliable predictions but also foster a more equitable income distribution among players, which could enhance team efficacy and cohesion. The evolution of salary prediction models has paralleled advancements in data analytics and machine learning, reflecting a growing recognition of the complexity of factors influencing player wages. Football salaries are shaped by multifaceted variables, ranging from individual performance metrics to broader economic trends and club revenues. Prior research has demonstrated a correlation between player performance and remuneration, serving as a basis for employing statistical models to improve prediction accuracy [5], [7]. Various algorithms and statistical methods are being explored to objectively analyze player worth, thus guiding clubs during negotiations in a data-driven manner.

Moreover, the financial viability of football clubs is heavily influenced by their payroll management. Accurate salary predictions can enable clubs to maintain sustainable wage bills while simultaneously optimizing their rosters. Unsustainable salary commitments could lead to financial crises, as evidenced by instances where clubs have faced insolvency due to mismanaged wage expenditures [8], [9]. Implementing robust salary prediction frameworks allows clubs to align player salaries with their revenue-generating capabilities, thereby enhancing both club viability and competitive balance across leagues. The contextual backdrop of the COVID-19 pandemic has further accentuated the importance of accurate salary prediction models, creating financial uncertainty and instability in the sports industry. The pandemic has fundamentally altered the revenue streams of many football clubs, emphasizing the necessity for clubs to reassess their financial frameworks [5], [10]. Establishing reliable salary prediction models in such turbulent times can help organizations make informed decisions regarding contract renewals and player acquisitions, ensuring that both clubs and players adapt to the evolving landscape of football economics.

The principal objective of utilizing machine learning for predicting soccer player salaries revolves around improving the accuracy and reliability of these forecasts while fostering a fair and more streamlined labor market within the world of professional sports. The complexity of football player salaries stems from various interlinked factors, including individual performance metrics, market conditions, club revenues, and broader economic indicators. Machine learning models, by leveraging large datasets, can uncover hidden patterns and interactions within these variables, consequently leading to more precise salary predictions [2]. Accurate salary prediction models have significant implications for clubs, players, and agents, enabling informed decision-making that enhances negotiation processes. The predictive power of advanced machine learning techniques can aid sports administrators in establishing equitable salary structures. By employing algorithms that account for numerous influencing factors, clubs can manage their payroll more effectively, aligning player salaries with revenue-generating capabilities, thus reducing the likelihood of financial distress. This is crucial in today's football landscape, where mismanagement of player wages can lead to severe financial ramifications, including insolvency.

The significance of this research lies in its potential to revolutionize football management, sponsorship, and team analysis by providing a data-driven approach to salary prediction. By accurately forecasting player salaries, teams can optimize their budgets, make informed decisions on player acquisitions, and ensure financial sustainability. Sponsorships can also benefit, as knowing the salary expectations of players helps brands target the right endorsements based on a player's marketability and influence. Additionally, the insights gained from analyzing player demographics

and team associations can support performance evaluations and improve strategies for team composition, ultimately enhancing overall team performance and competitiveness.

2. Literature Review

2.1. Previous Studies on Salary Prediction

Over the past few years, the application of machine learning models in the prediction of salaries, particularly within sports contexts, has gained considerable attention from researchers seeking to leverage data-driven methodologies for improved accuracy and fairness in remuneration. The existing literature on salary prediction emphasizes various models and techniques, pointing towards a dynamic landscape driven by advances in computational methods. One of the prominent studies is that by Bao [2], who emphasizes the potential of advanced machine learning models to enhance salary prediction accuracy through mathematical refinements and the integration of comprehensive datasets [2]. The study highlights how incorporating economic theory and statistical analyses not only improves model performance but also aids in understanding variable interactions that influence salary determinations. This is particularly critical in contexts where salary structures may be influenced by various interpersonal and market factors.

In a contrasting domain, research by Kim et al on predicting nurse turnover introduces the widespread use of the RF algorithm, which has proven effective in achieving high predictive accuracy an aspect also echoed in salary-related studies [11], [12]. The application of such robust algorithms has paved the way for analysts to harness similar strategies when addressing the complexities of player salary data. Further insights can be drawn from Papadaki and Tsagris [13], whose examination of NBA players' salaries indicates that more sophisticated machine learning algorithms such as Support Vector Machines (SVM) or gradient boosting can yield greater accuracy in predictions [13]. By exploring both player performance metrics and salary data, this research supports the idea that a nuanced approach one that considers a multitude of factors can lead to more precise estimations of player worth.

The studies by Chen et al. [14] reflect the diverse perspectives within salary prediction research. They explore the relationship between candidates' resumes and their expected salaries, implementing various algorithms such as multiple linear regression and decision trees to anchor their findings in a broader labor market context [14]. Simultaneously, a dual-Adaboosting system to predict salaries, addressing the limitations found in conventional regression models that often struggle with larger and more complex datasets [14]. Such innovations highlight the pressing need for advanced methodologies in accurately predicting salaries that encompass a variety of influencing factors outside of conventional performance indicators. In corporate settings, the trend to apply machine learning for salary predictions is also gaining traction. For instance, in the data science industry, Jiang et al. [15] exploration of salary trends through machine learning models illustrates a burgeoning interest in utilizing advanced analytical techniques to derive insights into salary determinations over time [15], [16]. This approach mirrors the trends seen in numerous domains including the sports sector where salary dynamics become increasingly intricate.

2.2. Random Forest Algorithm

RF is a powerful ensemble machine learning algorithm that has gained widespread acclaim for its efficacy in both classification and regression tasks across numerous domains, including sports analytics. Introduced by Sanjaykumar et al. [17], the method operates by constructing multiple decision trees during the training phase and outputs the mode of their predictions (for classification) or the average prediction (for regression). The inherent strengths of RF stem from its ability to manage large datasets with high dimensionality and complex interactions among variables, making it particularly well-suited for applications such as salary prediction in sports and other related performance analyses [18]. One of the most significant advantages of the RF algorithm is its robustness against overfitting, especially compared to single decision trees. This resilience is primarily due to the mechanism of averaging multiple trees, which reduces variance and prevents the model from becoming too closely tailored to a particular dataset [19]. This feature is crucial in contexts where data may be sparse or feature sets may exhibit high levels of noise. In practical applications, such as predicting player salaries, this quality allows for more reliable insights by mitigating the risk of over-general conclusions that may arise from singular decision-making models.

Another considerable advantage of RF lies in its ability to assess variable importance, providing insights into which features significantly influence the predictions. For instance, Silvino et al. [18] note that RF can deftly handle complex relationships in performance outcomes, allowing analysts to identify key performance indicators that drive results. This interpretability is particularly useful in sports analytics, wherein understanding the contribution of various metrics can aid in strategic planning for player development, training optimization, and even contract negotiations based on expected performance. In the context of salary prediction, studies have shown that employing RF can yield superior results compared to traditional methods. Al-Asadi and Taşdemir [6] demonstrated that the accuracy of salary predictions markedly improved when utilizing RF models to analyze FIFA video game data, illustrating the algorithm's ability to glean actionable insights from intricate player statistics. Similarly, the performance of RF has proven effective across multiple studies in sports analytics, affirming its adaptability for varying sports contexts, from soccer to cricket [20], [21].

2.3. Feature Engineering in Sports Analytics

In the realm of sports analytics, effective feature engineering plays a crucial role in developing predictive models that can accurately forecast player performance, salaries, and other significant outcomes. The utilization of player attributes in these models encompasses a wide array of metrics that capture both physical and technical-tactical indicators. These attributes serve as foundational elements driving the predictive capabilities of various machine learning algorithms, including classification and regression models. Feature engineering refers to the process of selecting, modifying, or creating new features from existing data to enhance model performance. In sports analytics, this can be particularly intricate due to the variety of metrics available for athletes. For instance, Adeyemo et al. [22] demonstrate the effectiveness of applying feature selection methods to optimize classification models for rugby league players by analyzing 157 physical and technical-tactical variables. Such comprehensive feature sets enable analysts to differentiate between players of various competitive levels, emphasizing the importance of meticulously crafted input data in developing robust analytical models.

Additionally, the integration of diverse player attributes reflects complex relationships between physical capabilities and performance outcomes. As noted by Ramos et al. [23], training experience, morphologic traits, and fitness attributes are significant predictors of performance in youth basketball players. This insight highlights the multifactorial nature of athletic success, where a combination of various attributes, including age, experience, and physical fitness, contributes to individual outcomes. By leveraging this complexity, predictive models can be fine-tuned to accurately assess each player's potential based on a multitude of interacting factors. The ability to analyze spatio-temporal data through modern tracking technologies further expands the scope of feature engineering in sports analytics. Mountifield [24] emphasizes the role of sophisticated statistical methods and quantitative models in enhancing athletic performance. As data acquisition systems have advanced, the volume and complexity of player performance data have increased significantly, allowing for a richer and more nuanced understanding of player attributes [24], [25]. Through the application of machine learning, organizations can create structured frameworks that not only enhance team performance but also provide a competitive advantage based on rigorous data analysis.

3. Methodology

3.1. Data Collection

The dataset used in this study contains essential information on football players, including their age, nationality, team, position, and salary details. It was loaded from a CSV file named player_salaries.csv, which consists of 684 entries and 7 columns. Upon loading the dataset, basic information was reviewed using df.info() to confirm the data types and ensure there were no missing values in the target variable Annual salary. Additionally, summary statistics were generated with df.describe() to examine the overall distribution of numerical features such as Age and Annual. This inspection ensured that the data was suitable for further analysis and modeling. An important step in the data collection process was confirming the cleanliness of the dataset. A missing values check (df.isnull().sum()) was performed to ensure that there were no gaps in the key variables that would impact model training. Based on the initial inspection, we verified that the dataset was complete and ready for preprocessing. The dataset’s features, including player

information and salary figures, were assumed to be clean for the purposes of this research, allowing us to proceed with the subsequent analysis and modeling steps without requiring imputation or cleaning.

3.2. Exploratory Data Analysis (EDA) & Visualization

EDA was conducted to better understand the data and uncover patterns that could guide our model design. Visualizations, such as histograms and scatter plots, were used to analyze the distribution of key variables like Age and Annual salary. The Age feature was plotted using a histogram to observe the spread of player ages, while the Annual salary distribution was visualized to check for any skewness or outliers. These visualizations provided an initial understanding of the data and highlighted any important relationships that may exist between player demographics and salary. Further visual analysis was conducted to explore the distribution of players by Nation and Team. Bar plots were created to show the top 15 nations and teams with the highest player counts, giving insights into the global distribution of players and the teams with the most representatives. Additionally, a scatter plot was used to explore the relationship between a player's age and their annual salary, providing a visual understanding of how these two variables correlate. All EDA plots were saved in the research_visuals directory for easy access and reference during further analysis.

3.3. Data Preprocessing

Data preprocessing is a critical step in preparing the dataset for modeling. The target variable, Annual salary, was separated from the feature set, which includes Nation, Position, Team, and Age. The dataset was then analyzed to identify categorical features, namely Nation and Position, which require encoding for use in machine learning algorithms. A key consideration during this stage was handling categorical variables through one-hot encoding, which was implemented using the OneHotEncoder from sklearn. This technique transforms categorical variables into numerical form, making them compatible with machine learning models. After preprocessing the categorical features, the dataset was split into training and testing sets using train_test_split with a 80-20 ratio, ensuring that 80% of the data was used for training and the remaining 20% for testing. The random state was set to 42 to ensure reproducibility of the results. The preprocessing pipeline, which included one-hot encoding of the categorical variables and maintaining the numerical feature Age unchanged, was established using ColumnTransformer. This pipeline ensures that categorical variables are properly handled while the remaining features are left intact for training.

3.4. Model Selection & Training

In this research, the RF Regressor was selected as the model for predicting football player salaries due to its flexibility and robustness in handling both categorical and numerical data. The RF algorithm is an ensemble method that uses multiple decision trees to improve prediction accuracy and reduce overfitting. The model was initialized with 100 estimators (n_estimators=100), which is a typical choice to balance performance and computation time. Additionally, the random_state=42 was used to ensure reproducibility of the model's results across different runs. The n_jobs=-1 parameter enabled the use of all available CPU cores, significantly speeding up the training process. The model was embedded in a Pipeline, where it was combined with the preprocessing steps. This pipeline first applied the ColumnTransformer to one-hot encode the categorical features and then used the RandomForestRegressor to fit the model to the data. The model was trained on the preprocessed training data (X_train, y_train), allowing it to learn from both player demographics and team associations. Training the model in a pipeline ensured a seamless workflow from data preprocessing to model fitting, minimizing the risk of data leakage and simplifying the evaluation process.

3.5. Model Evaluation

After training the model, it was evaluated on the testing set (X_test, y_test) to measure its performance. Standard regression metrics, including R-squared (R^2), Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE), were used to assess the model's accuracy. R^2 measures the proportion of the variance in the target variable that is explained by the model. A higher R^2 value indicates a better fit. MAE quantifies the average magnitude of errors in predictions, while RMSE provides a measure of the average magnitude of errors with a higher penalty for larger errors. These metrics provided a comprehensive understanding of the model's performance. The evaluation also included visual comparisons between the predicted and actual salaries. A scatter plot was generated to visually assess the closeness of the predicted salary values to the actual ones. The plot also included a reference line representing a perfect

 ISSN 2776-8007

prediction, allowing for easy comparison. This visual analysis supplemented the numerical evaluation and helped identify any systematic errors or patterns in the model's predictions.

3.6. Feature Importance Analysis

Feature importance analysis is a valuable aspect of the RF algorithm, as it helps identify which features most influence the model's predictions. The .feature_importances_ attribute of the trained RF model was used to extract the importance scores of each feature. These scores indicate the relative importance of each feature in predicting the target variable, Annual salary. The categorical features, such as Nation and Position, were first one-hot encoded, and their importance in salary prediction was assessed. The remaining features, such as Age, were also evaluated for their contribution to the model. A bar plot was created to visualize the top 20 most important features based on their importance scores. This visualization provided clear insights into which player attributes had the strongest influence on salary predictions. Understanding feature importance helps not only in model interpretation but also in refining the model by focusing on the most relevant features. This analysis could provide actionable insights for football teams and sponsors regarding the most influential factors in determining player salaries.

3.7. Model Checkpointing

Once the model was trained and evaluated, it was saved for future use or deployment using the joblib library. Saving the trained model ensures that the entire workflow, including preprocessing and model fitting, can be easily reused without the need to retrain. The model was saved as a file named random_forest_salary_predictor.joblib to the specified MODEL_SAVE_PATH. This step is crucial for any real-world application of the model, as it allows for efficient and consistent predictions on new data. By saving the model, the research ensures that the results can be easily accessed and applied in future football salary prediction tasks without the need for retraining. The checkpointing process also ensures that the model's performance and results can be replicated in the future. This is important for maintaining consistency in any deployed application, whether for research purposes or practical use in football management, sponsorship, or team analysis. By saving the model, the research ensures its longevity and usability beyond the immediate scope of the analysis.

4. Results and Discussion

4.1. Result

4.1.1. Results of Data Overview and EDA

The dataset used for this study consists of 684 entries and 7 columns, including features such as player name, nation, position, team, age, weekly salary, and annual salary. The dataset contains a mix of numerical and categorical variables. The average age of the players is approximately 24.76 years, with a standard deviation of 4.42 years. The Weekly and Annual salary columns show considerable variation, with the annual salary ranging from £27,923 to £25,402,050. Missing values were present in the Nation and Position columns, with 181 missing entries for Nation and 7 missing entries for Position. These missing values were handled during the preprocessing stage, ensuring no negative impact on model training.

EDA was performed to understand the distribution of the key numerical and categorical variables. The Age feature was visualized using a histogram (Figure 1), showing a roughly normal distribution with the majority of players aged between 21 and 27 years. The Annual salary distribution revealed a skewed distribution, with most players earning significantly less than the highest salaries, indicating the presence of outliers in the data.

 ISSN 2776-8007

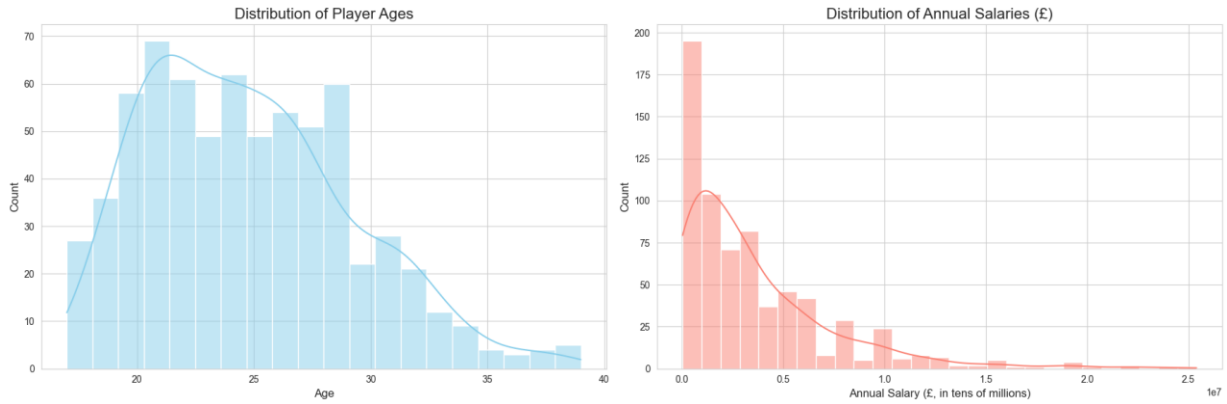

Figure 1. Histogram of Numerical Features

Categorical features, such as Nation and Team, were also analyzed. The bar plots of the top 15 nations and teams by player count (Figure 2) showed that certain countries and teams dominate the dataset in terms of player representation. These visualizations helped identify key patterns that could influence salary prediction and informed the model training process.

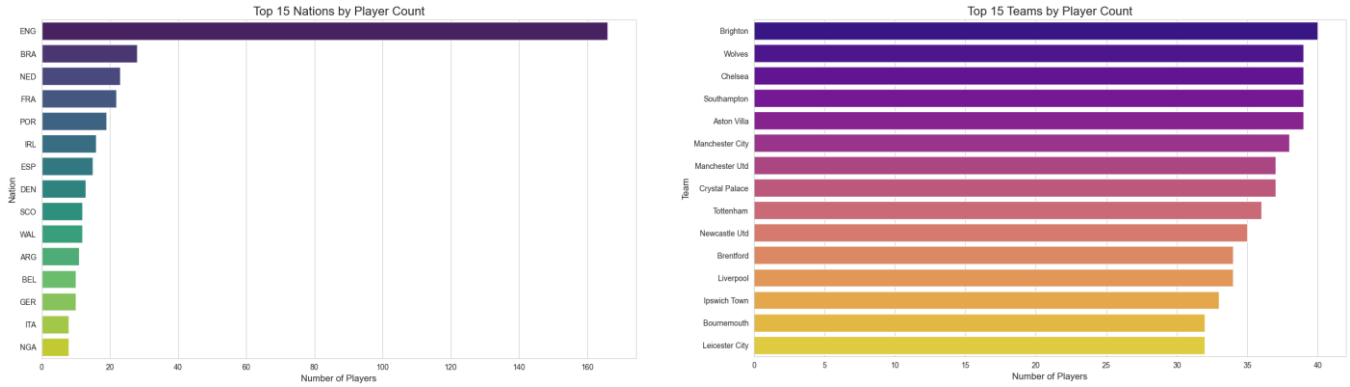

Figure 2. Bar Plot of Categorical Features

The dataset was preprocessed for model training by handling missing values and encoding categorical variables. The missing values in Nation and Position were identified and imputed during preprocessing, allowing the dataset to be used without any gaps. Categorical features, such as Nation, Position, and Team, were one-hot encoded using the OneHotEncoder to convert them into a numerical format suitable for machine learning models. The dataset was then split into 80% training data (547 samples) and 20% testing data (137 samples). This split ensured that the model could be trained on a large portion of the data while retaining a separate set for testing its generalization ability.

4.1.2. Results of Model Training and Elevation

The RF Regressor was selected as the model for predicting the annual salary of football players due to its robustness in handling non-linear relationships and its ability to manage both categorical and numerical data without requiring extensive feature scaling. The model was initialized with 100 estimators (n_estimators=100) and trained on the preprocessed training set. The use of the RF model, which combines multiple decision trees, allowed for more accurate predictions and helped mitigate overfitting. The model was trained successfully, and the training process was completed without any issues. The next step was to evaluate the model's performance on the testing set to assess how well it generalized to unseen data.

The trained model's performance was evaluated using several standard regression metrics. The R^2 value of 0.4853 indicates that the model explains approximately 48.53% of the variance in the annual salary data, which suggests a moderate level of predictive power. The MAE was £1,919,676.16, indicating that on average, the model's predictions were off by about £1.92 million. The RMSE of £2,823,642.83 further reflects the model's performance, with larger errors penalized more heavily. While the model performs moderately, there is room for improvement, particularly in

 ISSN 2776-8007

reducing the error rates for the more extreme salary predictions. A scatter plot of the actual versus predicted salaries was generated (Figure 3) to visually assess the model's predictions, highlighting the differences between the observed and predicted values. This visualization provided further insights into the areas where the model could be improved.

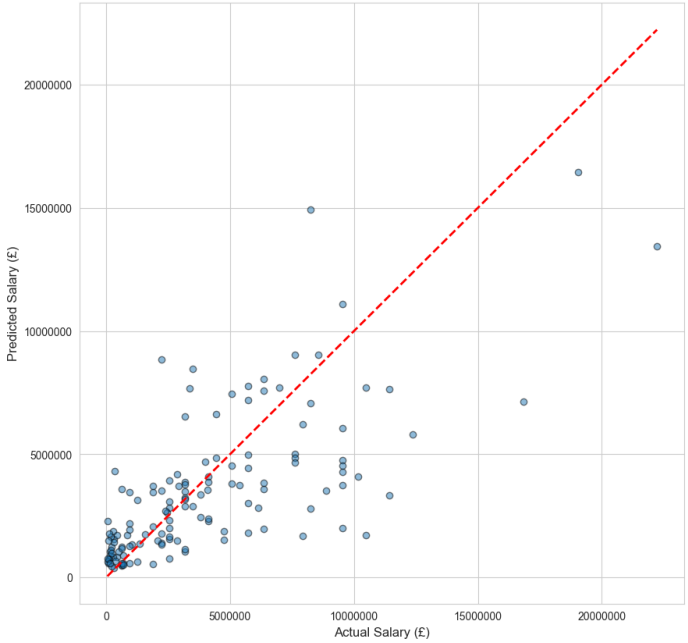

Figure 3. Actual vs Predicted Annual Salaries

4.2. Discussion

The RF model demonstrated a moderate performance in predicting football player salaries, as indicated by the R^2 value of 0.4853. This means that while the model is able to explain nearly 49% of the variance in player salaries, there is still a substantial amount of unexplained variance. Several factors may contribute to this. First, the dataset might not include all relevant features that influence salary, such as player performance metrics, marketability, or injury history. The model primarily relied on demographic and team-related features, which, although important, may not capture the full scope of factors that determine salary in the football industry. Adding more variables related to player performance or financial elements could potentially improve the model's predictive power.

The relatively high MAE of £1,919,676.16 suggests that the model's predictions are, on average, off by nearly £2 million, which is significant in the context of player salaries. This indicates that while the model is reasonable for rough estimates, it struggles to predict salaries with high precision, particularly for players with higher earnings. The presence of outliers in the salary distribution, where top players earn considerably more than the average, may skew the model's predictions. RF is less sensitive to outliers compared to other algorithms, but in this case, these extreme values still exert a noticeable influence. A potential solution could involve transforming the salary variable or applying a robust regression approach to better handle these outliers.

Despite these limitations, the model provides valuable insights into the relationship between player demographics and salary. The feature importance analysis revealed that categorical variables like Nation, Position, and Team had a substantial impact on the model's predictions. This suggests that factors such as the country of origin, team reputation, and position within the team significantly influence how salaries are set. However, the model's predictive accuracy could be enhanced by incorporating additional features, such as player performance statistics (goals, assists, etc.), league-specific economic factors, or even contract length. Future iterations of this model could benefit from a more comprehensive set of features and the exploration of other advanced algorithms, such as gradient boosting or neural networks, which might capture more complex patterns in the data.

5. Conclusion

This study found that the RF Regressor model, while offering a moderate level of predictive power, identified key predictors of football player salaries, such as player age, position, team, and nationality. The model explained about 48.5% of the variance in salaries, with categorical features like Nation and Team playing significant roles in the salary predictions. However, the model's accuracy was limited by the relatively high MAE, suggesting that more complex factors, such as player performance and external market influences, were not adequately captured. The findings have important practical implications for football team management and sponsorships. Understanding how factors like nationality and team influence salary could help teams make more informed decisions about player acquisitions, contract negotiations, and budget allocations. Sponsorships could also benefit by targeting players with higher marketability based on these attributes. However, the model's limitations, including its inability to fully account for outliers and missing performance metrics, suggest that future research could focus on incorporating more granular player data and testing advanced models like gradient boosting or deep learning to improve prediction accuracy.

6. Declarations

6.1. Author Contributions

Conceptualization: R.A.M.A., A.A.A.; Methodology: R.A.M.A., A.A.A.; Software: R.A.M.A.; Validation: A.A.A.; Formal Analysis: R.A.M.A.; Investigation: R.A.M.A.; Resources: A.A.A.; Data Curation: R.A.M.A.; Writing – Original Draft Preparation: R.A.M.A.; Writing – Review and Editing: R.A.M.A., A.A.A.; Visualization: R.A.M.A.; All authors have read and agreed to the published version of the manuscript.

6.2. Data Availability Statement

The data presented in this study are available on request from the corresponding author.

6.3. Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

6.4. Institutional Review Board Statement

Not applicable.

6.5. Informed Consent Statement

Not applicable.

6.6. Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] L. Yaldo and L. Shamir, “Computational Estimation of Football Player Wages,” *Int. J. Comput. Sci. Sport*, vol. 16, no. 1, pp. 18–38, 2017, doi: 10.1515/ijcss-2017-0002.

[2] Q. Bao, “Enhancing Salary Prediction Accuracy With Advanced Machine Learning Models,” *Appl. Comput. Eng.*, vol. 96, pp. 149-154, 2024, doi: 10.54254/2755-2721/96/20241185.

[3] Y. T. Matbouli and S. M. Alghamdi, “Statistical Machine Learning Regression Models for Salary Prediction Featuring Economy Wide Activities and Occupations,” *Information*, vol. 13, no. 10, p. 495, 2022, doi: 10.3390/info13100495.

[4] Z. Feng, Z. Liu, and Y. Yin, “Comparison of Deep-Learning and Conventional Machine Learning Algorithms for Salary Prediction,” *Appl. Comput. Eng.*, vol. 6, pp. 643-651, 2023, doi: 10.54254/2755-2721/6/20230910.

[5] T. K. Quansah, B. Frick, M. Lang, and K. Maguire, “The Importance of Club Revenues for Player Salaries and Transfer Expenses—How Does the Coronavirus Outbreak (COVID-19) Impact the English Premier League?,” *Sustainability*, vol. 13, no. 9, p. 5154, 2021, doi: 10.3390/su13095154.

[6] M. A. Al-Asadi and Ş. Taşdemir, "Predict the Value of Football Players Using FIFA Video Game Data and Machine Learning Techniques," *IEEE Access*, vol. 10, pp. 22631–22645, 2022, doi: 10.1109/access.2022.3154767.

[7] C. Thrane, "Performance and Actual Pay in Norwegian Soccer," *J. Sports Econ.*, vol. 20, no. 8, pp. 1051–1065, 2019, doi: 10.1177/1527002519851146.

[8] I. Perechuda, "Football Clubs Drowned by Players," *Polish J. Sport Tourism*, vol. 27, no. 1, pp. 28–32, 2020, doi: 10.2478/pjst-2020-0005.

[9] M. L. Doan, "Sentiment Trend Analysis of SpaceX Tweets Using Time-Series Sentiment Classification with TextBlob Algorithm," *J. Digit. Soc.*, vol. 1, no. 1, pp. 44-67, 2025, doi: 10.63913/jds.v1i1.3.

[10] I. G. A. K. Warmayana, Y. Yamashita, and N. Oka, "Analyzing the Impact of School Type on Student Outcomes Across Counties: A Comparative Study Using ANOVA," *Artif. Intell. Learn.*, vol. 1, no. 1, pp. 75-92, 2025, doi: 10.63913/jcl.v1i2.8.

[11] J. Kim, H. Kang, and P. Kang, "Time-series Anomaly Detection with Stacked Transformer Representations and 1D Convolutional Network," *Eng. Appl. Artif. Intell.*, vol. 120, p. 105964, 2023, doi: 10.1016/j.engappai.2023.105964.

[12] S. F. Pratama and A. M. Wahid, "Fraudulent Transaction Detection in Online Systems Using Random Forest and Gradient Boosting," *J. Cyber Law*, vol. 1, no. 1, pp. 88-115, 2025.

[13] I. Papadaki and M. Tsagris, "Are NBA Players' Salaries in Accordance with Their Performance on Court?," in *Contributions to Economics*, pp. 405-428, 2022, doi: 10.1007/978-3-030-85254-2_25.

[14] J. Chen, J. Hu, X. Xia, D. Lo, J. Grundy, Z. Gao, and T. Chen, "Angels or Demons: Investigating and Detecting Decentralized Financial Traps on Ethereum Smart Contracts," *Autom. Softw. Eng.*, vol. 31, p. 63, 2024, doi: 10.1007/s10515-024-00459-4.

[15] X. Jiang, Z. Dai, X. Pan, X. Lai, and J. Zhou, "A Review of Financial Services Research Based on Blockchain Technology," *Adv. Econ. Manag. Polit. Sci.*, vol. 92, no. 1, pp. 124–130, 2024, doi: 10.54254/2754-1169/92/20231231.

[16] A. M. Wahid, T. Hariguna, and G. Karyono, "Optimization of Recommender Systems for Image-Based Website Themes Using Transfer Learning," *J. Appl. Data Sci.*, vol. 6, no. 2, pp. 936-951, 2025, doi: 10.47738/jads.v6i2.671.

[17] S. Sanjaykumar, S. Natarajan, P. Y. Lakshmi, and F. A. Boby, "Predicting Team Success in the Indian Premier League Cricket 2024 Season Using Random Forest Analysis," *Phys. Educ. Theory Methodol.*, vol. 24, no. 2, pp. 304–309, 2024, doi: 10.17309/tmfv.2024.2.16.

[18] V. O. Silvino, L. G. da Silva Sousa, C. P. Ferreira, L. H. O. dos Santos, H. M. Apaza, S. S. Almeida, and M. A. P. dos Santos, "The Use of Machine Learning in Sports Performance: A Systematic Review," *Transl. J. Am. Coll. Sports Med.*, vol. 10, no. 2, e000304, Spring 2025, doi: 10.1249/TJX.0000000000000304.

[19] X. Jin, H. Q. Wang, and J. Zhong, "Anomaly Detection of Satellite Telemetry Data Based on Extended Dominant Sets Clustering," *J. Phys. Conf. Ser.*, vol. 2489, pp. 1-8, 2023, doi: 10.1088/1742-6596/2489/1/012036.

[20] C. Izumi and T. Hariguna, "In-Depth Analysis of Web3 Job Market: Insights from Blockchain and Cryptocurrency Employment Landscape," *Int. J. Res. Metaverse*, vol. 1, no. 1, pp. 40–58, 2024, doi: 10.47738/ijrm.v1i1.4.

[21] Y. Li, X. Yang, Q. Gao, H. Wang, J. Zhang, and T. Li, "Cross-Regional Fraud Detection via Continual Learning With Knowledge Transfer," *IEEE Trans. Knowl. Data Eng.*, vol. 36, no. 12, pp. 7865–7877, 2024, doi: 10.1109/TKDE.2024.3451161.

[22] V. E. Adeyemo, A. Palczewska, B. Jones, D. Weaving, and S. Whitehead, "Optimising Classification in Sport: A Replication Study Using Physical and Technical-Tactical Performance Indicators to Classify Competitive Levels in Rugby League Match-Play," *Sci. Med. Football*, vol. 7, no. 3, pp. 261–270, 2023, doi: 10.1080/24733938.2022.2146177.

[23] S. Ramos, A. Volossovitch, A. P. Ferreira, I. Fragoso, and L. M. Massuça, "Training Experience and Maturational, Morphological, and Fitness Attributes as Individual Performance Predictors in Male and Female Under-14 Portuguese Elite Basketball Players," *J. Strength Cond. Res.*, vol. 35, no. 7, pp. 2025–2032, 2021, doi: 10.1519/JSC.0000000000003042.

[24] C. Mountifield, "Sport Analytics: Graduating From Alchemy," in *Sport Analytics*, IntechOpen, 2023, doi: 10.5772/intechopen.1002423.

[25] T. Wahyuningsih and S. C. Chen, "Analyzing Sentiment Trends and Patterns in Bitcoin-Related Tweets Using TF-IDF Vectorization and K-Means Clustering," *J. Curr. Res. Blockchain*, vol. 1, no. 1, pp. 48–69, 2024, doi: 10.47738/jcrb.v1i1.11.