STA 663

Predicting Player Overall Ratings in FIFA 2022

Oluwabukola Emi-Johnson

Wake Forest University

12/14/2023

Table of Contents

Table of Figures

Abstract

Predicting overall outfield player ratings in FIFA 22 is challenging and is a hot topic amongst players of the video game. This paper explores prediction of the ratings using a dataset of 16,020 players and 45 key features. After meticulous data cleaning and exploratory data analysis, three predictive models—K-Nearest Neighbors (KNN), Elastic Net Regression, and Regression Trees—are employed and compared. KNN emerges as the most accurate, with a test RMSE of 1.84, outperforming Elastic Net Regression (RMSE: 2.07) and Regression Trees (RMSE: 2.42). The study recommends KNN for its superior predictive accuracy, emphasizing the importance of aligning model choice with data characteristics and analysis goals. The findings contribute valuable insights to the gaming community and highlight the application of machine learning in virtual sports analytics.

Section 1: Introduction

FIFA, the renowned soccer video game developed by EA Sports, has captured the enthusiasm of millions worldwide, boasting a player base exceeding 31 million on its FIFA 21 edition. Released annually, each new version of the game meticulously rates players, assigning them stats that aim to mirror their real-life counterparts. These stats encompass a spectrum of attributes, including crucial factors like Pace, Dribbling, and Shooting, all contributing to an overall rating.

In this paper, the primary objective is to predict the overall rating of outfield football players in FIFA 22, delving into the intriguing discussions among video gamers. The significance of this work goes beyond fueling discussions in the gaming community, it also aims to unravel the intricate factors influencing a player's virtual performance.

The analysis involves a robust dataset comprising 19,239 players, each defined by 110 features to develop a predictive model for the overall rating of outfield players, leveraging key attributes and providing valuable insights into the dynamics of virtual soccer performance. Some of the selected columns, including Age, Preferred Foot, Weight (kg), Height (cm), Pace (0-100), Shooting (0-100), Defending (0-100), Wage (in Euros), and Value (in Euros), constitute a high level of the diverse player characteristics embedded within the dataset. For a comprehensive understanding of all 110 columns, please refer to the dataset documentation available at FIFA 2022 complete player dataset.

Section 2: Data Cleaning and EDA

In the initial phase of data preparation, the focus was on identifying and extracting 46 key features crucial for predicting the overall ratings of outfield players. To align with the specific goal of targeting outfield players, exclusion criteria were applied to remove Goalkeepers and related player position information from the dataset.

Conducting thorough data quality checks revealed missing data for 1087 players. As a strategic decision, these instances were excluded from further analysis to uphold the integrity of the dataset. This action, considering the substantial dataset size, provided a cleaner and more robust dataset for subsequent analysis. Moreover, the categorical feature "preferred_foot" was transformed into a factor to enhance compatibility with the chosen analytical methods.

The meticulous data-cleaning process began with a dataset comprising 19,239 rows and 110 columns and resulted in a complete dataset of 16,020 rows and 45 columns. This dataset is now set for in-depth analysis, seeking to uncover the factors that influence the overall ratings of FIFA 22 outfield players.

In conducting Exploratory Data Analysis (EDA), the analysis focused on visualizing the response variable (Overall Ratings) alongside 10 key features, namely Preferred Foot, Age, Wage, Value, Pace, Shooting, Passing, Dribbling, Defending, and Physique. This comprehensive visual examination provides insights into the relationships and distributions within the dataset, paving the way for a more nuanced understanding of the factors influencing overall player ratings in FIFA 22.

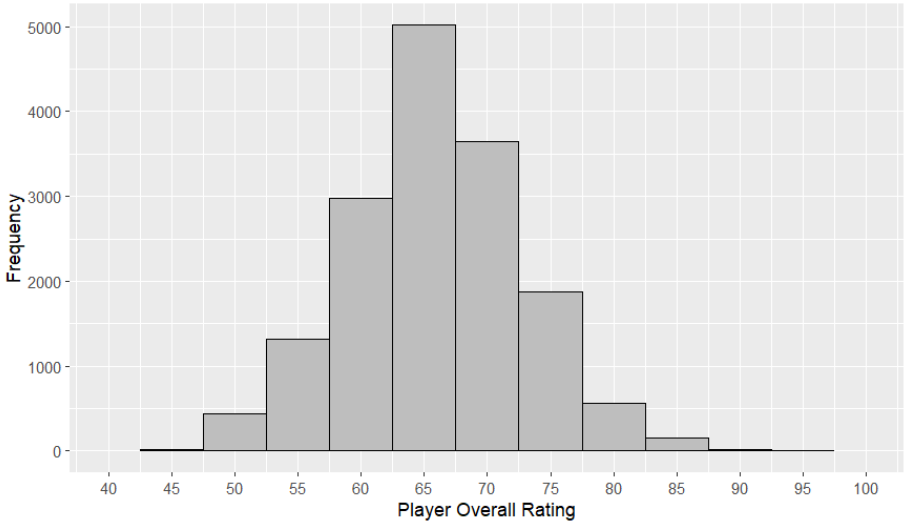

Figure 2.1: Distribution of FIFA 22 Overall Player Ratings

Figure 2.1 shows a histogram depicting the distribution of the response variable, Overall Player Ratings, across the 16,020 players in our dataset. The symmetrical shape of the histogram, centered around a rating of 65, indicates a normal distribution, with a substantial concentration of players around this central value. The spread extending from 47 to 93 shows the full range of player ratings, and highlights the difference in player abilities captured in the dataset. This visualization aids in understanding the central tendencies and variability within the overall player ratings.

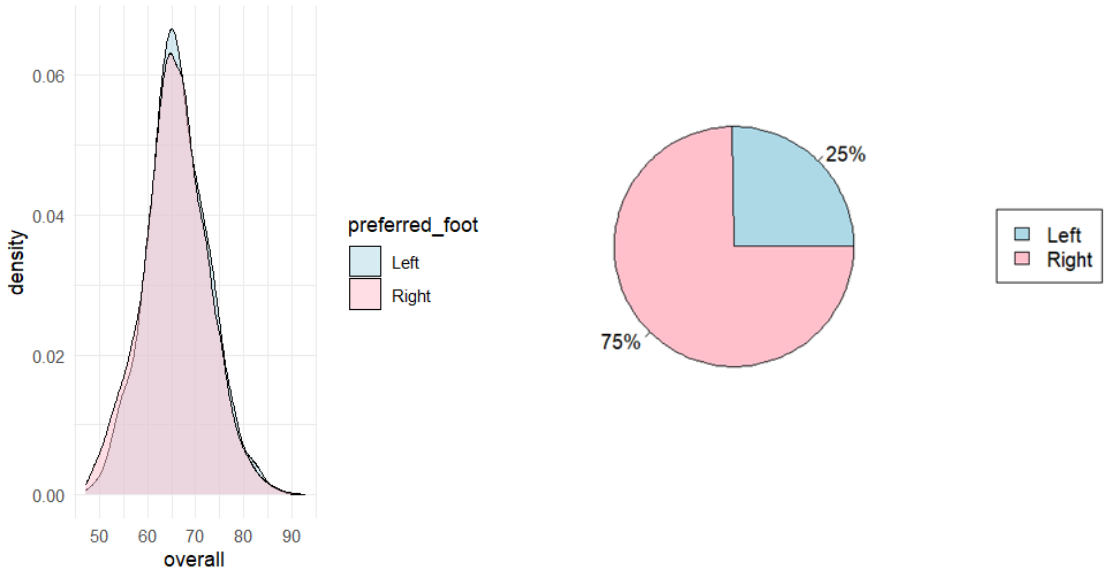

Figure 2.2: Distribution of Overall Rating by Preferred Foot and Proportions

Figure 2.2 illustrates that the Overall Ratings of both Left and Right footed Players, have similar distributions. Furthermore, the pie chart is a representation of the feature “preferred_foot” and indicates a clear preference for the right foot among players, with 75% favoring it over the left foot. This visualization effectively communicates the distribution of preferred foot orientations, emphasizing the dominance of right-footed players in our dataset.

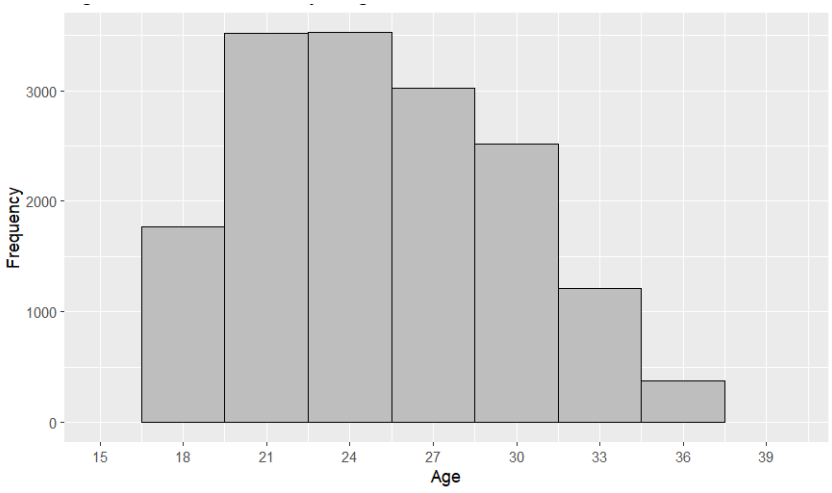

Figure 2.3: Distribution of Player Age

The histogram provides a high-level overview of the age distribution among players in our dataset, offering key insights into the demographic composition. Notably, the age range spans from a minimum of 16, capturing the youngest players in our dataset, to a maximum of 39, representing the oldest ones. Positioned at 25, the median age serves as a central reference point, indicating that half of the players are younger than 25, while the other half is older. This graphical representation effectively captures the diversity in age groups within our dataset, highlighting the various stages of players' careers.

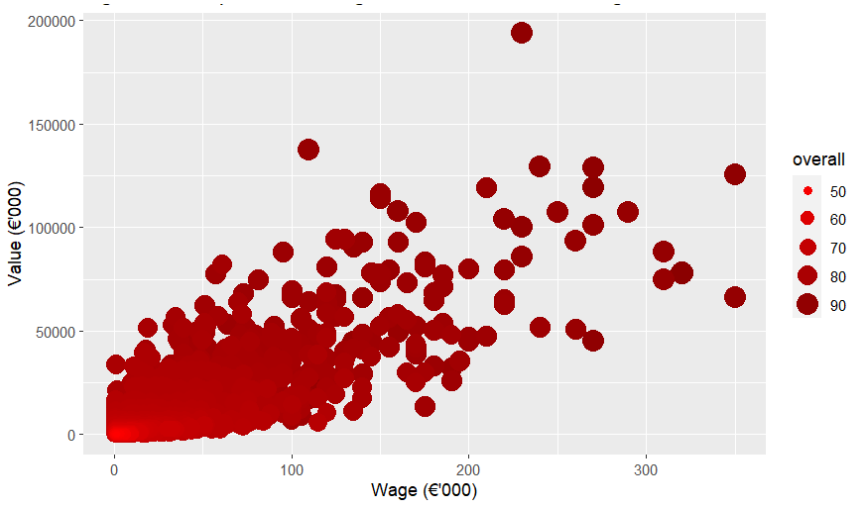

Figure 2.4: Comparison of Wage, Value and Overall Rating

Figure 2.4 provides a visualization of the relationship between Wage (in thousand euros), Value (in thousand euros), and Overall rating for football players. A relatively positive correlation is evident between the variables Wage and Value, indicating that players with higher wages tend to command higher values. This correlation underscores the close relationship between these two financial attributes in the context of football players. Furthermore, the plot introduces an additional layer of analysis by incorporating Overall rating as both the size and color parameter for each data point. The larger bubbles, denoting higher Overall ratings, are notably associated with players who command elevated wages and possess higher market values. This correlation aligns with the anticipated expectation that top-performing players, as reflected in their high Overall ratings, are not only valued more but also compensated with higher wages.

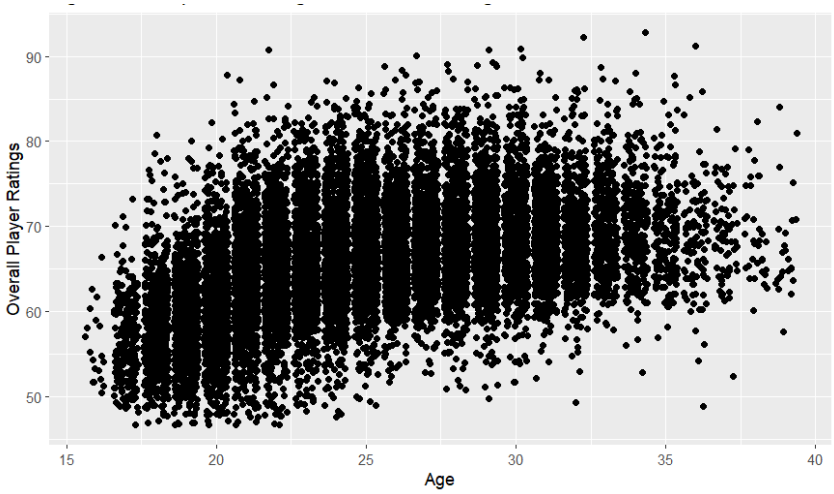

Figure 2.5: Comparison of Age and Overall Rating

Figure 2.5 provides a high-level representation of the relationship between Overall Player Ratings and Age. The graph unveils a distinct pattern, showcasing a positive trajectory in ratings for players aged 15 to 25, indicative of the pivotal developmental and formative stages in a football player's career. Within the age bracket of 25 to 32, the plateau in overall ratings suggests that players in this range have likely reached their peak performance levels. Finally, the decline in ratings observed among players aged 32 and over implies a natural decrease in performance levels as individuals

progress into the latter stages of their careers. This exploration of age-related variations in overall ratings contributes valuable insights into the dynamic nature of player performance across different age groups.

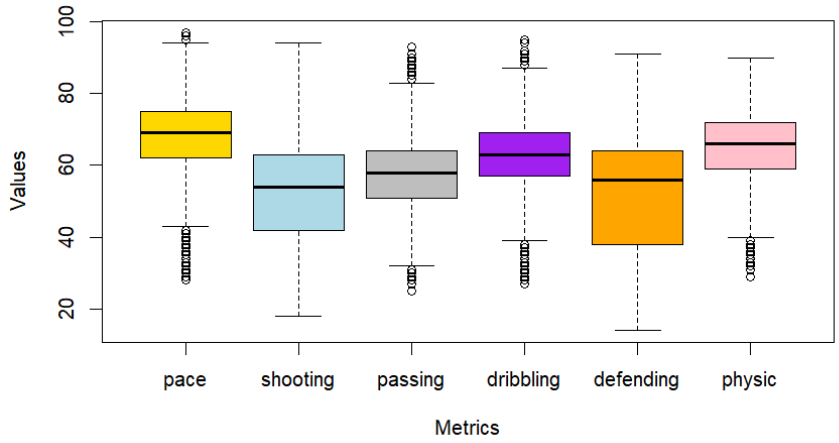

Figure 2.6: Distribution of Key Player Metrics

In Figure 2.6, an exploration of six key player attributes—Pace, Shooting, Passing, Dribbling, Defending, and Physique— is shown, shedding light on the diverse landscape of these characteristics among football players. A notable observation is the substantial range exhibited by Shooting and Defending, indicating a broader spectrum of variability in these attributes across the player population. Conversely, Physique presents the smallest range, signifying a more uniform distribution in this particular attribute. The distinct variability in attributes like Shooting and Defending may be indicative of varying playing styles and roles on the field, contributing to the strategic diversity within a team.

For the remaining attributes—Pace, Passing, Dribbling, and Physique—their smaller interquartile ranges (box length) suggest a more consistent distribution of values among the majority of players. However, the presence of outliers in these attributes highlights exceptional cases where certain players demonstrate either remarkably high or low performance levels. This analysis provides a high-level understanding of the intricate dynamics underlying the distribution of these key attributes, offering valuable insights into the diverse skill sets exhibited by football players.

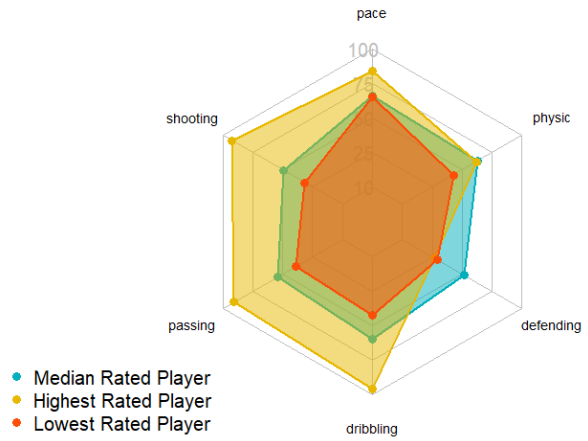

Figure 2.7: Comparison of Selected Players

For our final visualization for this section, the radar plot was used to compare player attributes and unveil insightful patterns among players with the Highest, Median, and Lowest overall ratings. A distinctive feature emerges as we scrutinize the Highest-rated player's plot, which exposes a discernible weakness in the Defending attribute. This observation hints at a probable offensive role for the player, emphasizing attacking prowess rather than defensive capabilities.

Contrastingly, when turning attention to the Median and Lowest-rated players, a notable similarity in the shapes of their radar plots becomes apparent. Although specific attribute values may differ between these two player groups, the relative distribution across attributes remains consistent. This consistent distribution implies that, despite variations in the absolute levels of individual attributes, the proportional composition of attributes contributes similarly to the overall ratings of players in both the Median and Lowest categories. This high-level analysis enhances our understanding of how different attributes collectively impact overall player ratings.

Section 3: KNN

Section 3.1: Introduction

In this section of the paper, we opt to employ the K-Nearest Neighbor (KNN) algorithm to predict Overall Outfield Player ratings in FIFA 22. The selection of KNN is well-suited for our analysis since it is capable of delivering highly accurate predictions, placing it in contention with some of the most precise models. Given that our primary focus is on achieving predictions with a high degree of accuracy, and we prioritize predictive performance over the interpretability of the model, KNN emerges as a good choice. The inherent flexibility and simplicity of the KNN algorithm align with our research objectives, making it a suitable and effective technique for predicting FIFA 22 Overall Outfield Player ratings.

Section 3.2: Method

The K-Nearest Neighbors (KNN) algorithm, used in this paper, operates as a machine learning algorithm used for both classification and regression tasks. The fundamental principle behind KNN lies in predicting outcomes based on similarity. To make predictions, the algorithm identifies the 'k' nearest data points in the training set, gauging proximity through a chosen distance metric. In our analysis, we opted for the Euclidean distance metric due to its effectiveness with continuous variables, its common application in practice, and its efficiency for our large dataset.

The parameter 'k' plays a pivotal role in shaping prediction accuracy, influencing the number of neighbors considered in the prediction process. To determine the optimal 'k' for our specific dataset, we systematically tested several values, including 3, 5, 10, 25, 75, and the square root of the dataset size (126) as is done in practice. Leveraging 10-fold Cross-Validation for robust evaluation, we calculated the test Root Mean Squared Error (RMSE) for each 'k' value. The resulting plot (Figure 3.1) showcases the relationship between 'k' and test RMSE, revealing a discernible trend. Through this analysis, we identified 'k = 5' as the optimal choice, aligning with the lowest test RMSE and thereby guiding our selection for the most effective 'k' in our predictive model.

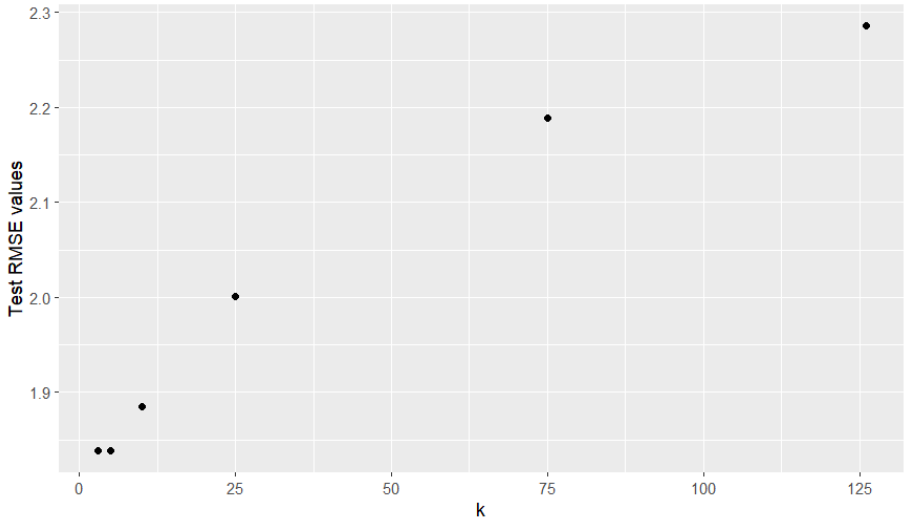

Figure 3.1: Visualization of Changes in RMSE with Changes in K

Section 3.3: Results

In obtaining the optimal 'k' value in our K-Nearest Neighbor (KNN) algorithm, we implemented a 10-fold cross-validation (CV) technique. The outcome of this process was a test Root Mean Squared Error (RMSE) value of 1.84. This metric serves as a significant measure of our model's predictive accuracy when utilizing the 5-Nearest Neighbors (5NN) approach. The test RMSE of 1.84 signifies a moderate yet commendable level of accuracy, implying that, on average, our model's predicted ratings deviate from the actual ratings by this value. This result highlights KNN algorithm's ability to capture intricate patterns within the dataset, showcasing its predictive ability.

However, it is important to acknowledge the inherent strengths and limitations of this approach. The algorithm excels in handling complex patterns but remains sensitive to the choice of the distance metric and the scale of the dataset. The need to convert a categorical feature into an indicator variable, as a result of our chosen distance metric, is one such consideration. Despite these shortcomings, the test RMSE reflects a pretty good accuracy level; a good indicator of KNN algorithm’s ability to predict Overall Player Ratings.

Section 4: Elastic Net Regression

Section 4.1: Introduction

In trying to enhance the accuracy of Overall Player rating predictions, we explore the use of Elastic Net regression in this section. Elastic Net is a hybrid regularization technique, combining L1 (Lasso) and L2 (Ridge) penalties to optimize linear regression models. This unique combination is particularly beneficial in addressing the challenge of multicollinearity, a common issue where predictors (features) in a dataset are correlated.

The distinctive feature of Elastic Net lies in its ability to produce more interpretable models. It achieves this by selectively omitting less important variables and spotlighting key features that exhibit a substantial relationship with the response variable. This process, known as shrinkage and selection, becomes especially valuable when dealing with datasets a large number of predictors. Elastic Net's proficiency in pinpointing and emphasizing the most critical variables differentiates it from the K-Nearest Neighbor (KNN) method employed in the previous section. While KNN excels in capturing intricate patterns in the data, Elastic Net introduces an interpretability dimension by emphasizing key predictors, making it an insightful option for our analysis.

Section 4.2: Method

Elastic Net, as a regularization technique, employs a dual tuning parameter system consisting of α and λ. α governs the trade-off between L1 (Lasso) and L2 (Ridge) penalties, influencing the degree of feature selection and coefficient shrinkage. The optimization objective of Elastic Net is to minimize the sum of squared residuals while simultaneously considering both the absolute values of the coefficients (L1) and their squares (L2).

The inclusion of L1 penalty allows Elastic Net to perform feature selection by effectively reducing some coefficients to exactly zero, indicating the exclusion of less impactful predictors. Simultaneously, the L2 penalty maintains a balance, preventing excessive emphasis on correlated predictors. The delicate interplay between these penalties facilitates the creation of a model that strikes a balance between model complexity and predictive accuracy.

To determine the optimal values for α and λ, the Elastic Net algorithm often employs cross-validation by systematically assessing different combinations of α and λ to identify the configuration that best aligns with the desired balance between complexity and accuracy. For our data using this method, we find our optimal tuning parameters to be $\alpha = 0.06$ and $\lambda = 0$. This means that for our data set, Elastic Net reduces to Ordinary Least Squares Regression.

Section 4.3: Results

In assessing the performance of Elastic Net regression for predicting Overall Player Ratings, a 10-fold cross-validation (CV) technique was employed, yielding a test Root Mean Squared Error (RMSE) value of 2.07. This metric provides a meaningful measure of the average deviation between the model's predicted ratings and the actual ratings, indicating a moderate level of accuracy in our predictions.

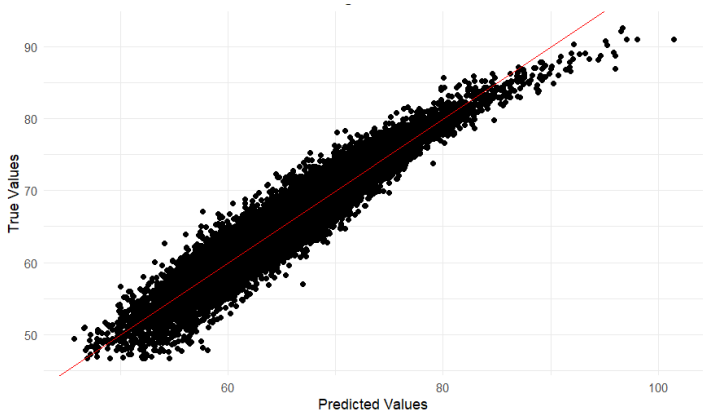

Figure 4.1: Plot of Predicted Ratings vs True Ratings

Figure 4.1 above shows the relationship between predicted overall ratings and their corresponding true values. A noticeable concentration of points along the diagonal line for predicted ratings up to 85 signifies relatively accurate predictions, demonstrating the model's reliability for a significant portion of the dataset. However, deviations become apparent for ratings above 85 highlighting areas where the model struggles to predict effectively.

The analysis shows good predictive performance, however, it is important to note a computational limitation tied to the method; solving for two tuning parameters (α and λ) which can be computationally expensive. Despite this limitation, the model showcases satisfactory predictive accuracy. It's important to highlight this trade-off between computational demands and predictive capability, particularly in scenarios where computational efficiency is a significant consideration.

In summary, the Elastic Net regression approach demonstrates moderate accuracy in predicting Overall Player Ratings.

Section 5: Regression Trees

Section 5.1: Introduction

In this section, our chosen method is Regression Trees, the third and final technique employed in our analysis. Regression Trees are a type of decision tree designed to accommodate independent variables with continuous values, offering a departure from the label-based leaves commonly associated with decision trees. The selection criteria and stopping criteria are also adapted to suit regression contexts.

The distinct advantage that Regression Trees bring to the analysis lies in their ability to provide a straightforward and intuitive representation of decision rules. This characteristic facilitates easier interpretation, particularly beneficial for non-technical audiences. In contrast to the Elastic Net regression used in the previous section, Regression Trees can handle complex and nonlinear relationships between predictors and the response variable whilst being less computationally expensive. This method adds interpretability and flexibility to the analysis, enabling a more in-depth exploration of the relationships within the dataset.

Section 5.2: Method

The process of growing a regression tree involves a recursive splitting of the dataset based on predictor variables, forming decision nodes that lead to terminal leaf nodes. The term "recursive" signifies the iterative nature of the process, where the dataset is subdivided into subsets using specified conditions or rules. At each decision node, a predictor and a split point are chosen to optimize a designated objective function; in this case, Residual Sum of Squares (RSS). This recursive splitting continues until a defined stopping criterion is satisfied, with our analysis setting a requirement for a 1% increase in the overall R-squared at each step. The leaf nodes represent predictions, calculated as the average of the Overall Player Ratings for the players within each leaf.

The resulting tree is interpretable, providing clear insights into the relationships between predictors and the target variable. However, precautions are taken to prevent overfitting, with techniques like pruning considered to enhance model performance. This ensures that the regression tree strikes a balance between capturing intricate relationships and maintaining generalizability to new data.

Section 5.3: Results

Following the analysis, the generated Regression tree (Figure 5.1) comprises 10 leaves resulting from 9 splits on variables such as Value, Age, and movement_reactions. Predictions derived from the tree were employed to estimate the test Root Mean Squared Error (RMSE) using 10-fold Cross-Validation (CV), resulting in a value of 2.42. This metric serves as a valuable measure of how the model's predictions deviate from the actual Overall Player Ratings, with a test RMSE of 2.42 indicating a moderate level of accuracy. On average, the predicted ratings differ from the actual ratings by this value.

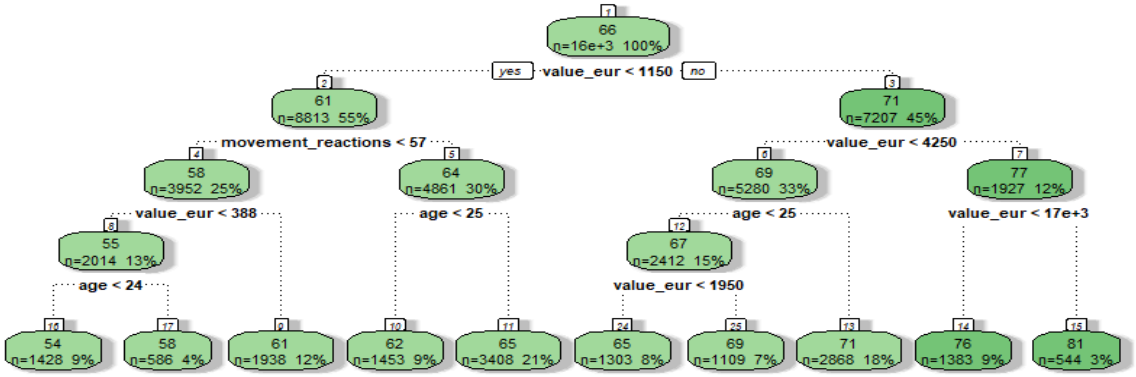

Figure 5.1: Regression Tree

Similar to the previous methods, Regression trees exhibit certain limitations. They may be prone to overfitting, especially when the tree is deep and captures noise in the data. Additionally, the instability of trees means that minor alterations in the dataset can lead to substantially different tree structures. Despite these drawbacks, the interpretability of regression trees and their ability to capture complex, nonlinear relationships between predictors and the response variable make them a valuable tool.

Conclusion

In conclusion of our analysis, it is evident that K-Nearest Neighbors (KNN) outperforms both Elastic Net Regression and Regression trees, achieving the lowest test Root Mean Squared Error (RMSE) of 1.84. This implies that, on average, our predicted overall player ratings deviate from the true ratings by approximately 1.84. The significantly lower RMSE of KNN signifies higher predictive accuracy compared to the other two methods under consideration, making it the superior choice for our analysis.

When selecting a model, it is crucial to consider the data's structure, features, and analysis goals. Our primary objective is to make precise predictions. Given KNN's superior predictive accuracy over Elastic Net and Regression Trees, it aligns with our objective. Therefore, its use is strongly recommended for predicting overall player ratings in this context.

Finally, it is important to note that KNN's effectiveness depends on the choice of 'k.' Fine-tuning this parameter might further enhance the model's accuracy. Also, regular validation and adjustment of the model, as the dataset evolves or expands, will ensure the continued reliability of our predictions.

Citation

FIFA 2022. Retrieved December 6th, 2023 from https://sports-statistics.com