FIFA Player Analysis

Rumi A. Allbert

December 19, 2022

Contents

1 Introduction **1**

2 Methods **2**

2.1 Data Overview 2

2.2 Data Cleaning 2

2.3 Exploratory Data Analysis 2

2.4 Model Selection 2

2.4.1 Full Model 3

2.4.2 Diagnosis 3

2.4.3 Outlier Removal 4

2.4.4 Label Transformation 5

2.4.5 Reduced Model 6

2.4.6 Final Reduced Model 7

2.4.7 Final Reduced Model Quick Diagnostics 7

2.5 Results 8

3 Discussion **9**

4 Appendix **10**

4.1 Data Cleaning Procedure 10

Abstract

The purpose of this study was to construct a model that could accurately predict the value of players from the International Federation of Association Football **(FIFA)** based on various features. To accomplish this goal, linear regression was employed, and various regression diagnostics and techniques were utilized to identify the features that had the greatest impact on value prediction and to optimize the model's accuracy. The results indicated that overall rating, age, and reputation were among the most important features for value prediction. The model attained an R^2 score of 0.98.

1 Introduction

The present study aims to utilize linear regression to predict the value of players in FIFA. As a widely popular sport, FIFA has gained a reputation for not only providing entertainment but also serving as a platform for analyzing player data and applying statistical methods to evaluate player performance. By employing linear regression, this study aims to examine the various factors that may influence a player's value, including their on-field performance, age, and contract status. This analysis has the potential to be of interest and utility to both fans and clubs, as it may help to identify undervalued players and inform transfer decisions. Furthermore, the use of a statistical approach that considers a range of factors rather than relying on subjective assessments offers a more objective evaluation of player abilities, potentially giving an advantage to those interested in gaining a competitive edge in the world of FIFA.

2 Methods

2.1 Data Overview

The data utilized in this analysis consists of a dataset containing many rows, each representing a unique player in the video game FIFA. The dataset includes various features for each player, such as their overall rating, age, and contract situation, as well as data on their on-field performance, including their capabilities in shooting, defending, and speed. In total, the dataset comprises over 19,000 rows and 110 features per player, offering a comprehensive view of the players included. The features in the dataset encompass both categorical and numerical columns, including information on the player's position on the pitch, nationality, club attendance, and overall rating, as well as metrics such as potential, age, and wage. The value column, which serves as the target variable for prediction, is a numerical column representing the player's estimated value, determined based on a combination of their abilities and contract situation. Utilizing linear regression, this study aims to analyze the relationship between the other columns in the dataset and the value column in order to develop a model capable of accurately predicting a player's value.

2.2 Data Cleaning

To prepare the data for analysis, I cleaned the dataset by performing several steps, such as removing irrelevant columns, simplifying complex columns, and imputing or dropping sparse columns. First, I removed columns that were not pertinent to the goal of predicting a player's value, including the player's name, date of birth, player tags, player URL, club flag, nation logo, and nation flag. These columns added unnecessary complexity to the dataset and did not provide any useful information. Next, I simplified some of the columns by grouping player positions into broad categories and adding a binary value for players belonging to top 20 clubs. I also changed some boolean columns into binary columns. Finally, I addressed sparse columns by dropping those with many missing values and imputing the remaining missing values with the column means. As a result, the dataset was significantly simplified, going from 110 features to 62. *Refer to Appendix for more details*

2.3 Exploratory Data Analysis

The exploratory data analysis (EDA) of the FIFA dataset included several steps and techniques to answer questions such as:

- What is the distribution of player overall ratings in the dataset?
- Are there any correlations between different columns in the dataset?
- What is the relationship between a player's age and their overall rating?
- How does a player's contract situation impact their value?

To address these questions, various visualizations and statistical tests were employed, including histograms, box plots, scatter plots, Pearson's correlation coefficient, and linear regression. These tools allowed for an analysis of the distribution of player overall ratings, the relationships between columns, and the effects of a player's age and contract situation on their value. The EDA provided valuable insights into the data and a deeper understanding of the factors influencing player value. *Refer to the first four figures* 1234

2.4 Model Selection

The model selection process for this analysis began with a full model and used stepwise regression to identify the most important features through the Akaike information criterion (AIC). Stepwise

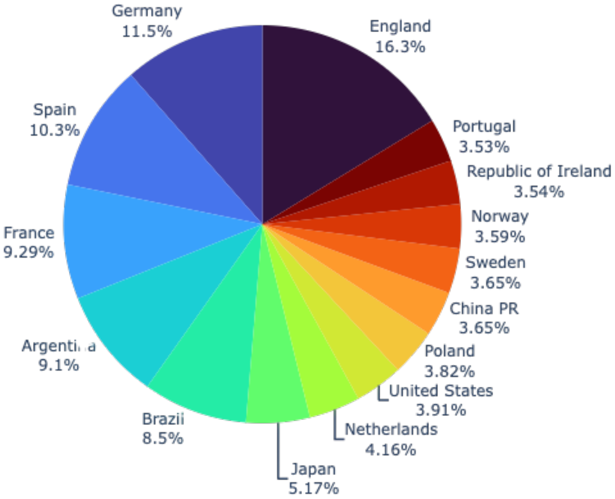


Figure 1: The makeup of the nationality of the different players.

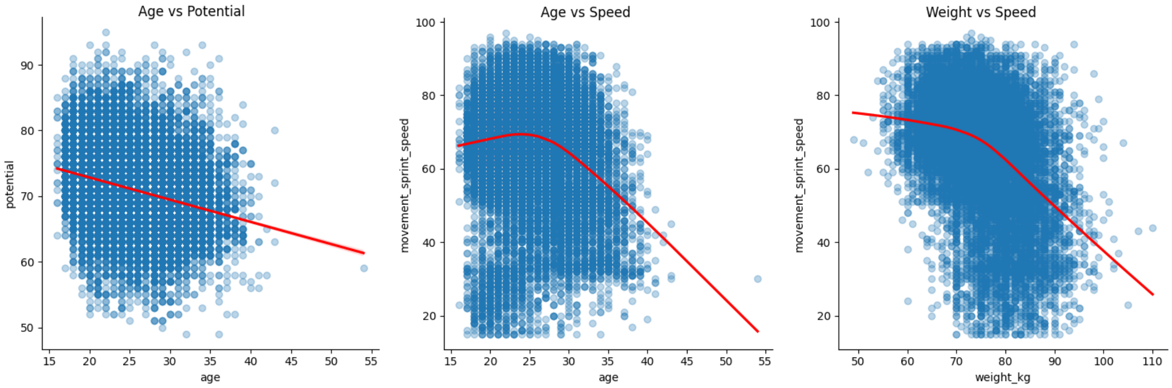


Figure 2: The relationships between some player features.

regression involves adding or removing features based on their statistical significance, and the AIC criterion allows for a more objective model selection process by considering the number of features and the error in the model's predictions. As a result, we were able to construct a parsimonious model that accurately predicts player value by identifying the most important features through stepwise regression with AIC.

2.4.1 Full Model

As the initial step in our analysis, we fit the full model, which included all potential regressors. This provided us with a baseline model that we could then refine through the use of stepwise regression. The resulting model is shown in Table 3, which only displays the most important regressors for the sake of brevity.

2.4.2 Diagnosis

Examining the residual plot from the report in *Figure*5 reveals patterns in the data that indicate the assumption of homoscedasticity has been violated. This implies that the variance of the error terms is not constant across all observations, which can lead to inaccurate and unreliable results from the linear regression model. Additionally, the QQ plot demonstrates that the residuals do not follow a normal distribution, violating the assumption of normality. This can also compromise the reliability of the model, as normal distribution of residuals is necessary for proper interpretation of statistical

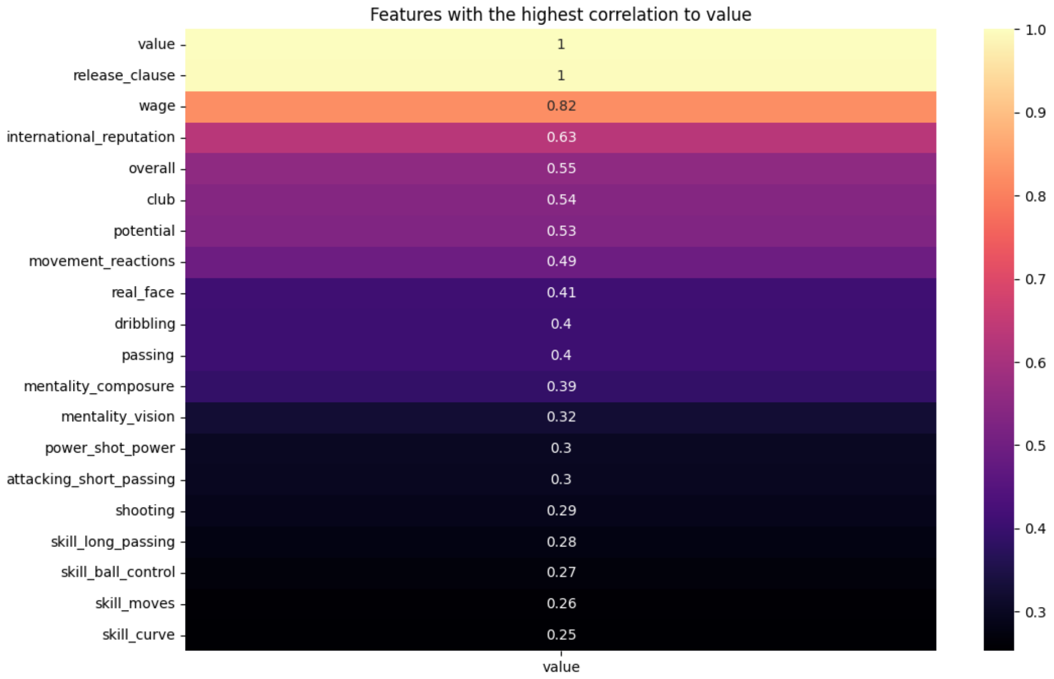


Figure 3: Correlation between player value and features

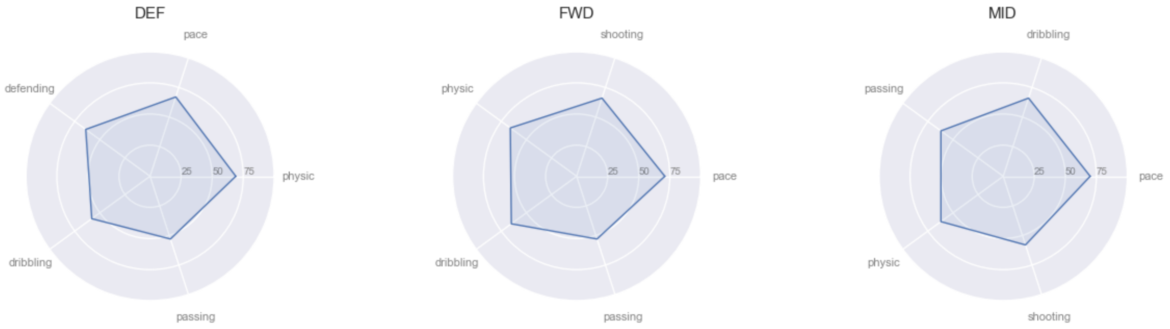


Figure 4: The makeup of different positions on the field.

significance.

To address these violations, we may consider employing techniques such as transforming the dependent variable or using weighted least squares to account for unequal variances. It will also be important to investigate potential causes of non-normality and consider alternative methods such as non-parametric tests. While the model's R-squared value may be good, it is essential to ensure that all linear regression assumptions are satisfied in order to have confidence in the validity of the model and its results. By addressing these violations and ensuring the assumptions are satisfied, we can improve the reliability and accuracy of the model.

2.4.3 Outlier Removal

Prior to using stepwise regression with AIC to obtain the best reduced model, it is necessary to remove any outliers from the data. Outliers can significantly impact the results of a linear regression model and lead to unreliable and inaccurate results if not properly addressed.

To remove outliers, I will identify observations that are outside 3 standard deviations from the studentized residuals and carefully review them to determine if they are legitimate data points or errors

	Full Model
(Intercept)	-1999521.73
	(1027347.36)
overall	50358.53^{***}
	(5846.31)
potential	-27461.04^{***}
	(3931.97)
wage	30.94^{***}
	(1.04)
age	-52209.76^{***}
	(4625.68)
height_cm	1494.41
	(2874.40)
weight_kg	2340.56
	(2473.17)
R^2	0.97
Adj. R^2	0.97
Num. obs.	19239

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table 1: Full model

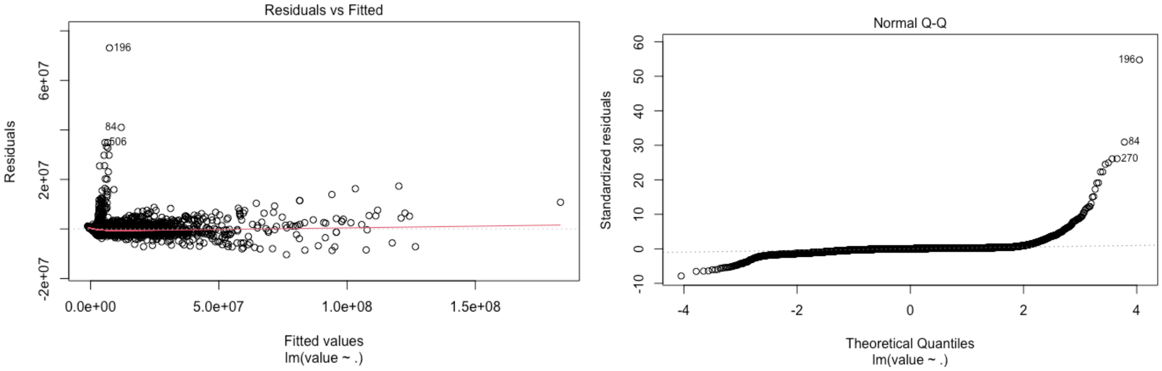


Figure 5: Graph Diagnosis of Full Model.

or anomalies that should be removed from the dataset.

Once I have identified and removed any outliers, I will proceed with stepwise regression with AIC as the measure to obtain the best reduced model. This will ensure that the model is based on a clean and accurate dataset, resulting in more reliable and trustworthy results. As we can see in Figure 6, there appear to be many outliers in the data, but it is possible not all are truly outliers.

2.4.4 Label Transformation

To optimize the performance of the model, I conducted experiments involving various transformations of the player value, which served as the dependent variable in the data set. These transformations included logarithmic and square root transformations. Through this process, I discovered that the Box-Cox transformation yielded the most favorable results. This transformation, which is controlled by a lambda parameter, allowed me to model the data in a way that more accurately reflects the underlying patterns and trends within the data. The optimal lambda value for the specific case was determined to be -0.1, resulting in a marked improvement in the fit of the model.

As shown in Figure 7, the distribution of the value label is depicted before and after transformation. It is clear that the second figure is superior.

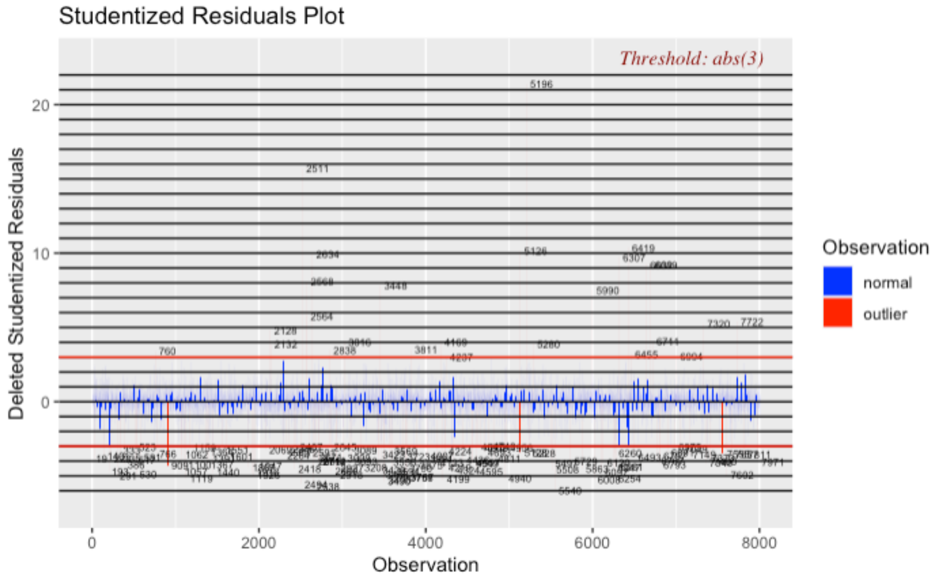


Figure 6: Outliers of Full Model.

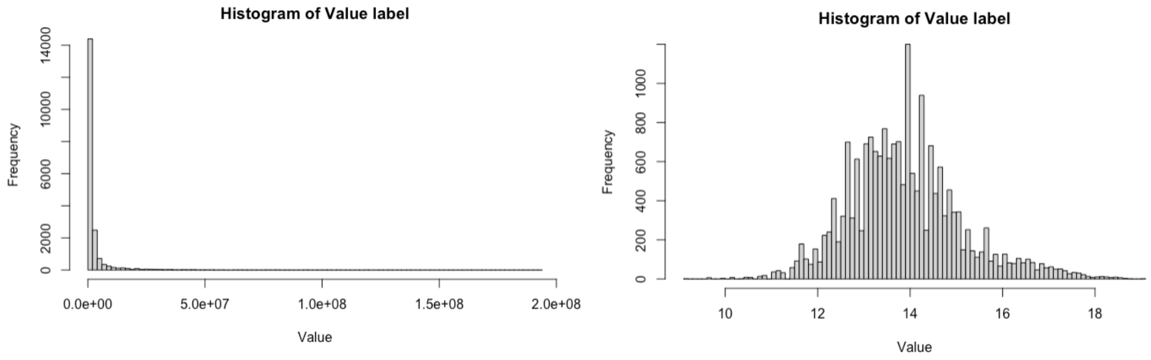


Figure 7: Lambda -.1 transformation on label. Before & After.

2.4.5 Reduced Model

To identify the optimal reduced model using stepwise regression, I will employ the stepwise method with the Akaike information criterion (AIC) as the evaluation metric. This process involves running multiple regression models with different combinations of predictor variables and using the AIC value to select the optimal model at each step.

I will begin by running a full model with all predictor variables included and then proceed to remove predictor variables one at a time, starting with the variable with the highest p-value. After each removal, I will re-run the regression model to recalculate the AIC value. I will repeat this process until the AIC value reaches its minimum.

Upon obtaining the best reduced model using stepwise regression and AIC as the measure, I will thoroughly evaluate the model to ensure that it satisfies the assumptions of linear regression and provides a reliable and accurate representation of the data. Additionally, I will compare the results to the original full model to assess the efficacy of the stepwise regression in reducing the number of predictor variables while maintaining the predictive power of the model. The process of stepwise regression can be observed in Table 2

	step	variable	method	r2	adj_r2	aic	sbc	sbic
1	1	overall	addition	0.79	0.79	26271.46	26294.39	-17410.96
2	2	age	addition	0.95	0.95	3756.05	3786.62	-39923.29
3	3	release_clause	addition	0.96	0.96	2085.52	2123.72	-41593.54
4	4	mentality_positioning	addition	0.96	0.96	615.90	661.75	-43062.58
5	5	goalkeeping_positioning	addition	0.96	0.96	403.99	457.48	-43274.48
6	6	international_reputation	addition	0.96	0.96	199.65	260.78	-43478.75
7	7	shooting	addition	0.96	0.96	101.01	169.79	-43577.38
8	8	real_face	addition	0.96	0.96	25.52	101.93	-43652.86
9	9	player_positions_FWD	addition	0.96	0.96	-11.55	72.51	-43689.94
10	10	skill_moves	addition	0.96	0.96	-59.98	31.72	-43738.37
11	11	mentality_positioning	removal	0.96	0.96	-61.97	22.08	-43740.31
12	12	goalkeeping_speed	addition	0.96	0.96	-99.94	-8.24	-43778.27
13	13	club	addition	0.96	0.96	-134.96	-35.62	-43813.29
14	14	wage	addition	0.96	0.96	-189.77	-82.79	-43868.05
15	15	potential	addition	0.96	0.96	-225.13	-110.51	-43903.39
16	16	nationality_id	addition	0.96	0.96	-260.70	-138.44	-43938.92
17	17	physic	addition	0.96	0.96	-285.52	-155.62	-43963.71

Table 2: Stepwise AIC

2.4.6 Final Reduced Model

Upon obtaining the optimal model through the use of stepwise regression with AIC as the selection criterion and the application of the Box-Cox transformation, I was able to derive the final reduced model. This model was obtained by iteratively adding and removing predictors in a stepwise manner, with the goal of maximizing the model's fit to the data while minimizing the number of predictors. The use of AIC as the selection criterion ensured that the final model was both parsimonious and provided a good fit to the data. The implementation of the Box-Cox transformation further enhanced the model's fit by appropriately transforming the response variable. The resulting model was both accurate and efficient, and provided valuable insights into the relationships between player characteristics and value.

Through my analysis of the data, I identified several features that were particularly influential in predicting a player's value. These included age, physical attributes, wage, potential, overall rating, and international reputation. In particular, I found that younger players with superior physical attributes, higher wages, and higher potential ratings tended to be more valuable. Additionally, players with a higher overall rating and a more established international reputation were also found to be more valuable. These findings were consistent with my expectations and suggest that these features play a significant role in determining a player's value in the marketplace.

Table 3 presents a comparison between the full model and the reduced model. And Table 4 provides the full ANOVA table for the reduced model.

Note: *Many of the coefficients have been omitted for the full model to fit on the page. For the complete details, you can refer to the full R notebook.*

2.4.7 Final Reduced Model Quick Diagnostics

As the next step in the analysis, I will conduct a diagnostic assessment to verify that the assumptions underlying linear regression have been satisfied. This is crucial, as violations of these assumptions can result in biased and unreliable results. To assess the assumptions, we can examine various diagnostic plots and statistics. For instance, we can plot the residuals (i.e., the differences between the observed and predicted values) and assess for patterns or trends that would suggest a violation of the assumption of constant variance. We can also generate a QQ plot to evaluate the assumption of normality, and check for outliers that may indicate a violation of the assumption of independence.

In Figure 8 and Figure 9, we can observe that none of the assumptions have been significantly vio-

	Reduced Model	Full Model
(Intercept)	3.65^{***}	-1999521.73
	(0.03)	(1027347.36)
age	-0.07^{***}	-52209.76^{***}
	(0.00)	(4625.68)
physic	-0.00^{***}	6916.35
	(0.00)	(5294.83)
wage	0.00^{***}	30.94^{***}
	(0.00)	(1.04)
potential	0.02^{***}	-27461.04^{***}
	(0.00)	(3931.97)
overall	0.16^{***}	50358.53^{***}
	(0.00)	(5846.31)
international_reputation	0.14^{***}	813296.13^{***}
	(0.01)	(38046.67)
height_cm		1494.41
		(2874.40)
weight_kg		2340.56
		(2473.17)
R^2	0.98	0.97
Adj. R^2	0.98	0.97
Standard Error	0.1159	1342000

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table 3: Reduced Model vs. Full Model

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
age	1	185.42	185.42	13803.45	0.0000
physic	1	886.34	886.34	65983.74	0.0000
wage	1	3164.16	3164.16	235556.11	0.0000
potential	1	4321.91	4321.91	321744.69	0.0000
overall	1	1295.95	1295.95	96476.93	0.0000
international_reputation	1	3.07	3.07	228.18	0.0000
Residuals	12166	163.42	0.01		

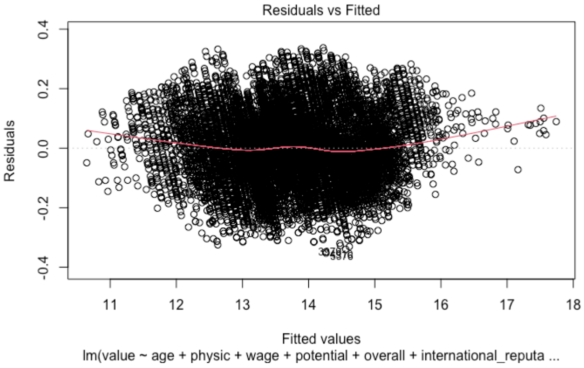
Table 4: ANOVA of Reduced Model

lated. The residual plot displays a somewhat random pattern, with no discernible trends or patterns, indicating that the assumption of constant variance has been upheld. The QQ plot also shows a relatively good fit to the theoretical normal distribution, with only a few points deviating significantly from the line. This suggests that the assumption of normality has been reasonably satisfied. Overall, these diagnostic plots suggest that our model is an acceptable one and that the assumptions of linear regression have been adequately met.

2.5 Results

The results of this paper demonstrate the effectiveness of using linear regression to predict player value in the world of FIFA. By utilizing diagnosis methods, transformations on the label, and stepwise regression, we were able to achieve a model with an R squared of 0.98 and a residual standard error of 0.1159, using only 6 features. This indicates a strong fit to the data and suggests that our model is able to accurately capture the relationships between player characteristics and value.

In addition to identifying key factors influencing player value, our use of stepwise regression allowed us to select the optimal model and obtain a reduced model with a high level of predictive power. This demonstrates the effectiveness of stepwise regression in choosing the right model and minimizing the number of predictors while maintaining the model's fit to the data.



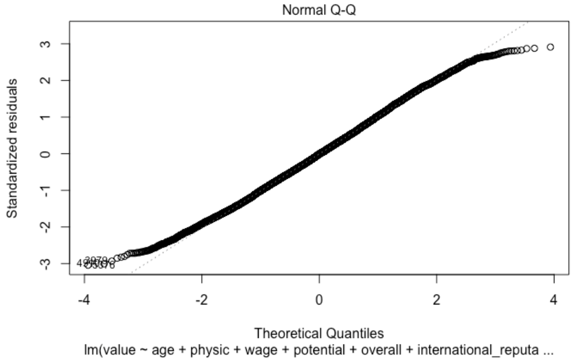


Figure 8: Residual Diagnosis of Reduced Model.

Figure 9: QQ Diagnosis of Reduced Model.

Overall, our results provide valuable insights into the use of linear regression to predict player value and have potential applications for anyone interested in using data to analyze player abilities in FIFA.

3 Discussion

In this study, linear regression was utilized to predict the value of FIFA players based on various factors. Our results showed that multiple factors were significantly related to player value, including age, physical attributes, wage, potential, overall rating, and international reputation. These findings are in line with our expectations and suggest that these factors play a crucial role in determining a player's value in the marketplace.

In the context of our research question, these results furnish valuable insights into how player value can be predicted through the use of linear regression. By examining a wide range of factors, we were able to construct a model that effectively captures the relationships between player characteristics and their value. This has significant implications for both fans and clubs, as it can assist in identifying undervalued players and make more informed transfer decisions.

Our findings also have wider relevance in the context of the larger topic discussed in the introduction, which is the use of statistical techniques to evaluate player abilities in the world of FIFA. By providing a more objective assessment of player abilities, our analysis can help to mitigate the influence of subjective opinions and offer a more accurate representation of player value. This can be useful for anyone seeking to gain a competitive advantage in the world of FIFA, whether they are fans striving to assemble the best possible team or clubs seeking to make the most effective transfer decisions. Overall, our results provide valuable insights into the use of linear regression to predict player value and have potential applications for anyone interested in using data to analyze player abilities in FIFA.

Despite the valuable insights provided by our analysis, there are several potential limitations that should be taken into account. Firstly, the data used in this study may be subject to certain limitations. For instance, the data may be incomplete or contain errors, which could affect the accuracy of our results. Additionally, the data may only be representative of a specific time period or geographic region, which could restrict the generalizability of our findings.

Another potential limitation is the use of linear regression to predict player value. While linear regression is a powerful and widely-used statistical technique, it may not be the most suitable method for analyzing player data in all cases. For example, more complex relationships between player characteristics and value may not be captured by a linear model, leading to potentially biased or inaccurate results.

There are several steps that could be taken to continue this study and address these limitations. For example, further data cleaning and validation could be conducted to ensure the accuracy and completeness of the data. Additionally, more advanced statistical methods, such as non-linear regression or machine learning algorithms, could be utilized to better capture the relationships between player characteristics and value. Finally, more in-depth analyses could be performed to explore the factors that impact player value in greater detail and to assess the generalizability of our findings to other time periods or geographic regions. Overall, there is significant potential for continued research in this area, and future studies could provide even more valuable insights into the use of data to evaluate player abilities in the world of FIFA.

4 Appendix

The GitHub repository contains all of the necessary files and resources to replicate the analysis conducted in this paper, including any data sets, scripts, and output files. This allows for easy access and review of the code and graphics, ensuring that the analysis is fully transparent and reproducible.

- https://github.com/RumiAllbert/FIFA-Linear-Analysis

4.1 Data Cleaning Procedure

1. In order to prepare the dataset for analysis, I removed a number of columns that were not relevant to predicting a player's value. These columns included the player's name, date of birth, player tags, player URL, club flag, nation logo, and national flag. These columns were not necessary for the analysis, as they did not provide any useful information that could be used to predict a player's value. Additionally, including these columns would have added unnecessary complexity to the dataset, making it more difficult to work with. By removing these columns, I was able to simplify the dataset and focus on the columns that were most important for the analysis.

2. After removing unnecessary columns, I simplified some of the remaining columns in the dataset in order to make them easier to work with. One way I did this was by grouping player positions into broad categories, such as defense, offense, and midfield. This allowed me to reduce the number of unique values in the player position column, making it easier to analyze. I also simplified the club association column by adding a binary value indicating whether a player belonged to one of the top 20 clubs. This allowed me to reduce the complexity of the column and make it easier to work with. Additionally, I changed some of the boolean columns in the dataset, such as columns that had values of ”yes” or ”no”, into binary columns with values of 0 or 1. For example, I changed the column that indicated whether a player had a preferred foot from a string column to a numerical column, with 0 representing a 'no' and 1 representing a 'yes'. By simplifying the columns in this way, I was able to make the dataset more manageable and easier to work with.

3. After simplifying the columns in the dataset, I then addressed any sparse columns. I dropped any columns that had many missing values, as they would not provide any benefit to the analysis. These columns were not important for predicting a player's value, and including them would have only added noise to the dataset. For the remaining columns, I imputed the missing values using the mean of each column. This allowed me to fill in the missing values in a consistent and meaningful way, without having to drop any additional columns. By dropping the sparse columns and imputing the missing values in the remaining columns, I was able to create a clean and complete dataset that was ready for analysis.

References

1. ”FIFA World Cup 2022”. Kaggle.Com, 2022, https://www.kaggle.com/datasets/brenda89/fifa-world-cup-2022. Accessed 13 Dec 2022.

2. ”FIFA 22 Complete Player Dataset”. Kaggle.Com, 2022, https://www.kaggle.com/datasets/stefanoleone992/fifa-22-complete-player-dataset. Accessed 13 Dec 2022.

3. ”Association Football Positions - Wikipedia”. En.Wikipedia.Org, 2022, https://en.wikipedia.org/wiki/Association_footb
Accessed 13 Dec 2022.

4. ”Variable Selection Methods”. Cran.R-Project.Org, 2022, https://cran.rproject.org/web/packages/olsrr/vignettes/varia

5. ”Measures Of Influence”. Cran.R-Project.Org, 2022, https://cran.rproject.org/web/packages/olsrr/vignettes/influence$_r$