Assessing Premier League Player Valuation: The Impact of Age, Position, Playtime and Other Variables on Market Value

Stat Warriors

Introduction

We all have an interest in soccer and we were curious about how players were valued compared to their teammates. Across numerous leagues, soccer players have been valued primarily based on their performance, which can be broken down into many factors such as age, position, and minutes played, alongside other key factors. We believe that these three will be the most crucial factors in determining a player's valuation, leading us to formulate the following questions: 1) Do older players have a higher transfer market value than younger players, on average? 2) Do offensive players have a higher transfer market value than defensive players? 3) Do players with a higher play time (minutes played) have a greater transfer market value than players with a lower play time?

We would like to study whether our hypothesis is true, and if any other relevant factors are as influential, if not more. Examining the rationale behind transfer market values and players' performances helps clubs make smarter decisions about their players. For example, younger players, particularly those under 21, often garner higher fees since they are seen as investments for the future (Metelski, 2021). Additionally, examining player positions is crucial, as forwards usually command higher fees (Zhou, 2019). Minutes played can also significantly impact a player's value as it enhances their performance metrics and leads to increased player visibility. (Besson, Ravenel, Poli, 2020).

Methods and Analysis

Initially the histogram of the response was strongly skewed and not continuous, so we started by transforming our response variable. After applying a log transformation, our histogram of response was approximately unimodal, symmetric and continuous.

Then, we performed the exploratory data analysis (EDA). Age looked to have a negative, linear correlation, minutes played overall looked to have a positive, linear correlation, and the means on the box plots for all three qualitative variables (position, club ranking group, and England) were different across each level.

After performing our EDA, we checked for multicollinearity within our quantitative variables. The average VIF was above 3 and there are multiple individual VIFs above 10. Since we had concerns for multicollinearity, we continued with variable screening in an attempt to resolve these issues. Using stepwise regression with a p-ent and p-rem of 0.15, our new qualitative model included the variables: clean sheets overall, age, goals involved per 90 overall, and minutes per match. Our variable screening removed: minutes played overall, appearances overall, goals overall, assists overall, and conceded overall. Afterwards, we checked for multicollinearity again, as we no longer had concerns about it since the average VIF was below 3 and each individual VIF was below 10.

Next, we started adding in our qualitative predictors one at a time. We first added position and then performed a nested F-test to determine if position type is significant. Our p-value was larger than 0.05, so position was not significant at predicting valuation. Thus, we removed all 3 dummy variables relating to position. Next, we added club ranking group and performed another nested F-test. This determined that club ranking group is significant, so we kept all 3 dummy variables and did not test them individually. Finally, we added the England variable and performed a nest F-test to determine its significance. We found player nationality to not be significant, so we removed it from the model.

The next step of our analysis was to test for interactions. With the remaining quantitative variables and the qualitative variable of club ranking groups, we checked to see if any interactions were present. After viewing the interaction plots between the various combinations, we saw that there were no interactions present, so we continued with our analysis.

Our updated model at this point of the analysis included clean sheets overall, age, goals involved per 90 overall, minutes per match, and club ranking group. Next, we tested our regression assumptions. The lack of fit assumption was not violated because, when looking at the residual plot, there was no odd trend in the data, as all residuals had a mean of zero. When looking at the Residuals vs Fitted plot, the residuals were spread evenly across the predicted values and across the x values, indicating no violation of the constant variance assumption. Looking at the Q-Q residual plot and the histogram of residuals, we verified that the data was not radically skewed, so the normality assumption was not violated. We do not have time series data, so the independence assumption was not violated.

The last part of our analysis is the influential diagnostics. After analyzing Cook's distance, leverage (hat), studentized residuals, and deleted studentized residuals, we found observations 51 and 85 (Gerard Deulofeu and Stuart Armstrong, respectively) to be outliers in the x direction, 24 and 38 (Anthony Martial and Theo Walcott, respectively) as outliers in the y direction, and 24, 25, and 38 to be influential observations. We decided to remove all outliers and influential observations, ultimately refitting our model based on the updated model.

Our final model after all of our testing and methods included clean sheets overall, age, goals involved per 90 overall, minutes per match, and club ranking group. We used cross validation as our added technique to verify how capable our model is at predicting with new data. Our findings will be explained in the results section below.

Results

The model shows all predictors have a VIF below 2, indicating no significant multicollinearity. The parameters for the key variables are statistically significant ($p<0.05$). The final model was validated using a 5-fold cross-validation (external model validation), with R-Squared of 0.7982, root mean square error of 0.4594 and a mean absolute error (MAE) of 0.3611. Hence, the model is significant, indicating accurate and consistent predictions across subsets. 79.82% of the variance in the log of valuation is explained by the predictors. Ultimately, we believe that our final model we developed is a good model for determining the transfer value of the soccer players.

Conclusions

Our final prediction equation is log(Valuation) = 4.706709+0.044463(CleanSheetsOverall) - 0.14508(Age) + 1.457279(GoalsInvolvedPer90Overall) + 0.019431(MinPerMatch) + 0.935701(ClubRankingOneFive) + 0.547181(ClubRankingSixTen) - 0.243145(ClubRankingSixteenTwenty). Our model is useful for predicting player valuations based on our external validation. It is applicable for clubs to estimate transfer fees based on measurable and categorical attributes. For our example, we used Anthony Martial's data, a Manchester United striker from the '18-'19 Premier League season. After inputting his relevant values into the model, we got an approximate log valuation of 3.84536. We saw relatively consistent results with the cross-validation, which, again, indicated a fairly high predictive capability of our model.

We faced various limitations in our analysis such as being limited to 5 or fewer categorical levels. This led us to group the 20 premier league clubs into groups of 4 according to their league position at the end of the season. Moreover, as we only analyzed data for one premier league season, we are not able to cross reference how player transfer value is changing over time. We had to also filter our data to reduce multicollinearity as in sport related data, there tend to be many variables that overlap with one another.

In the future, we could add more variables into our analysis, such as passes made per 90, shots taken per 90, and the injury record of players. We could also expand our analysis across multiple seasons, allowing us to incorporate inflation into our transfer market value. Lastly, we could manipulate the data to change the weightage of variables according to player position (Ex: Clean sheets would have a higher impact on the valuation of a defensive player than an attacking player).

Appendix A: Data Dictionary

Variable Name	Abbreviated Name	Description
Valuation	val	Valuation of player in millions of dollars
logval	logval	log of the Valuation of player in millions of dollars
Age	age	Age of the player in years
Position	pos	Position of player (forward, defender, goalkeeper)
Club.Ranking	rank	Rank of club as categories 1-5, 6-10, 10-15, 16-20
minutes_played_overall	min	Total number of minutes played throughout the entire season
England	eng	Whether or not the player is from England
appearances_overall	app	Number of times the player appears on the field
goals_overall	goal	Total number of goals the player scored throughout the season
assists_overall	ast	Total number of assists the player had throughout the season
clean_sheets_overall	clean	Total number of clean sheets the player had throughout the season. A clean sheet is when the team prevents the other team from scoring any goals in the match
conceded_overall	conc	Total number of conceded games the player had throughout the season. A conceded game is when the team fails to stop the opposing team from scoring points or goals
goal_involved_per_90_overall	invo	The number of goals a player would score if they played a full 90 minute match, based on their current scoring rate
min_per_match	minper	The average minutes a player plays in a 90 minute match

Appendix B: Data Rows

	Player.Name	Age	Position	Current.Club	Club.Ranking.Group
1	Ederson	29	Goalkeeper	Manchester City	One_Five
2	Kyle Walker	32	Defender	Manchester City	One_Five
3	Raheem Sterling	27	Forward	Manchester City	One_Five
4	Vincent Kompany	36	Defender	Manchester City	One_Five
5	Alisson Becker	30	Goalkeeper	Liverpool	One_Five
6	Andrew Robertson	28	Defender	Liverpool	One_Five

	Club.Rankings	Valuation	minutes_played_overall	minutes_played_home
1	1	70	3420	1710
2	1	50	2777	1245
3	1	140	2777	1256
4	1	8	1223	561
5	2	80	3420	1710
6	2	60	3219	1599

	minutes_played_away	nationality	England	appearances_overall	appearances_home
1	1710	Brazil	No	38	19
2	1532	England	Yes	33	15
3	1521	England	Yes	34	16
4	662	Belgium	No	17	8
5	1710	Brazil	No	38	19
6	1620	Scotland	No	36	18

	appearances_away	goals_overall	goals_home	goals_away	assists_overall
1	19	0	0	0	1
2	18	1	1	0	1
3	18	17	12	5	10
4	9	1	1	0	0
5	19	0	0	0	0
6	18	0	0	0	11

	assists_home	assists_away	penalty_goals	penalty_misses	clean_sheets_overall
1	1	0	0	0	20
2	0	1	0	0	18
3	6	4	0	0	18
4	0	0	0	0	7
5	0	0	0	0	21

6	9	2	0	0	21

	clean_sheets_home	clean_sheets_away	conceded_overall	conceded_home
1	9	11	23	12
2	7	11	17	9
3	7	11	19	9
4	3	4	7	3
5	12	9	19	8
6	12	9	17	7

	conceded_away	yellow_cards_overall	red_cards_overall
1	11	2	0
2	8	3	0
3	10	3	0
4	4	6	0
5	11	1	0
6	10	4	0

	goals_involved_per_90_overall	assists_per_90_overall	goals_per_90_overall
1	0.03	0.03	0.00
2	0.06	0.03	0.03
3	0.88	0.32	0.55
4	0.07	0.00	0.07
5	0.00	0.00	0.00
6	0.31	0.31	0.00

	goals_per_90_home	goals_per_90_away	min_per_goal_overall
1	0.00	0.0	0
2	0.07	0.0	2777
3	0.86	0.3	163
4	0.16	0.0	1223
5	0.00	0.0	0
6	0.00	0.0	0

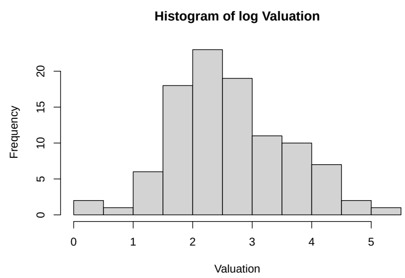
	conceded_per_90_overall	min_per_conceded_overall	min_per_match
1	0.61	149	90
2	0.55	163	84
3	0.62	146	82
4	0.52	175	72
5	0.50	180	90
6	0.48	189	89

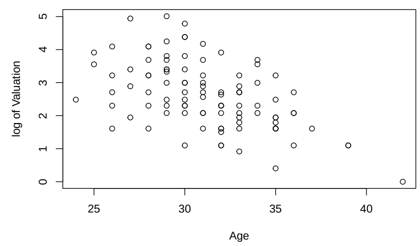
	min_per_card_overall	min_per_assist_overall	cards_per_90_overall
1	1710	3420	0.05
2	926	2777	0.10
3	926	278	0.10
4	204	0	0.44
5	3420	0	0.03
6	805	293	0.11

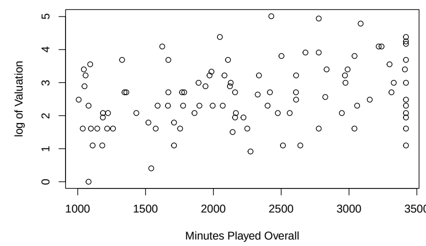
	rank_in_league_top_attackers	rank_in_league_top_midfielders
1	366	240
2	250	248
3	12	14
4	194	384
5	417	387
6	418	19

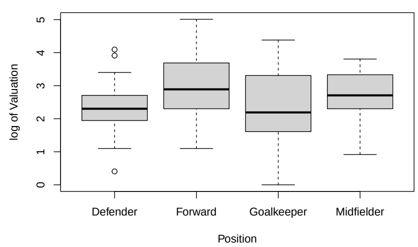
	rank_in_league_top_defenders	rank_in_club_top_scorer
1	13	21
2	10	14
3	-1	2
4	7	12
5	3	16
6	2	20

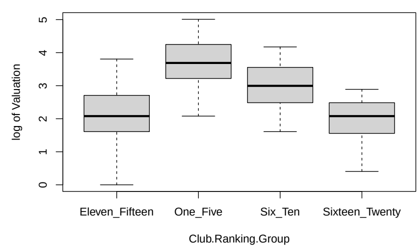
Appendix C: Tables and Figures

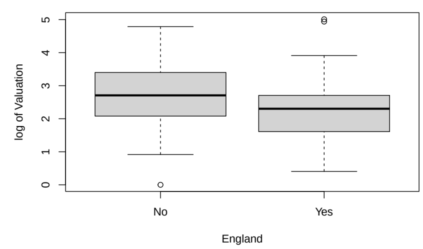


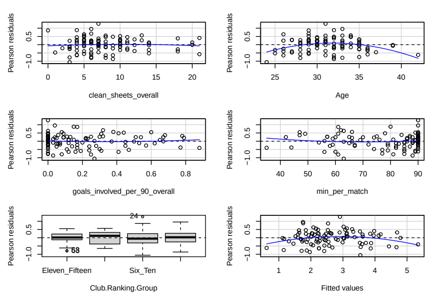


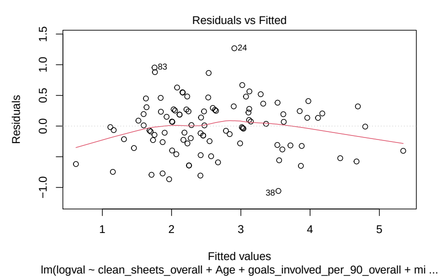


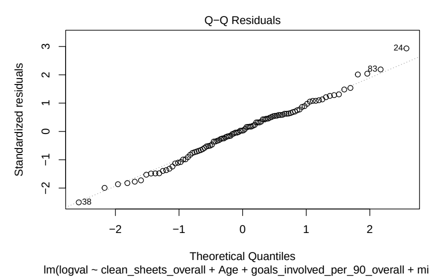


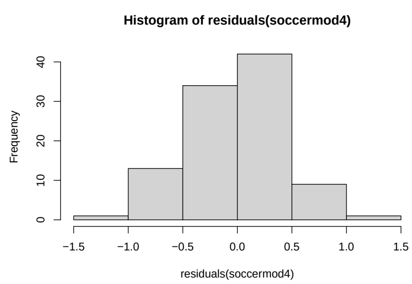


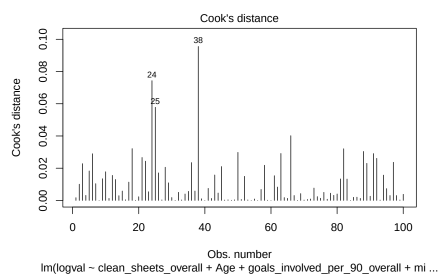


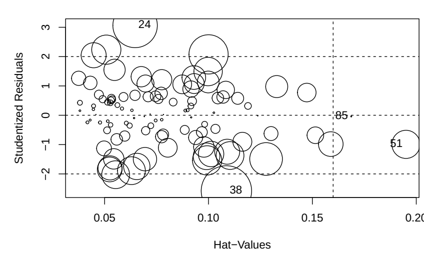




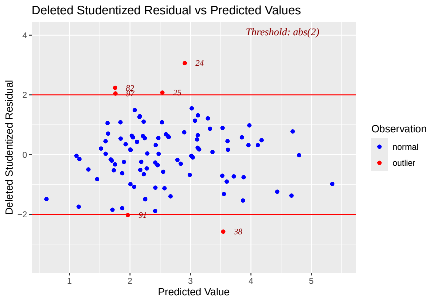








	StudRes	Hat	CookD
24	3.06322779	0.06467541	7.433126e-02
38	-2.58039151	0.10862247	9.554707e-02
51	-0.99225306	0.19504893	2.982647e-02
85	-0.04127371	0.16871391	4.369138e-05




```
           Valuation                          Age
                   1                            0
            Position           Club.Ranking.Group
                   0                            0
minutes_played_overall                    England
                   0                            0
 appearances_overall                goals_overall
                   0                            0
     assists_overall         clean_sheets_overall
                   0                            0
    conceded_overall goals_involved_per_90_overall
                   0                            0
       min_per_match                       logval
                   0                            1
```

```
           Valuation                          Age
                   0                            0
            Position           Club.Ranking.Group
                   0                            0
minutes_played_overall                    England
                   0                            0
 appearances_overall                goals_overall
                   0                            0
     assists_overall         clean_sheets_overall
                   0                            0
    conceded_overall goals_involved_per_90_overall
                   0                            0
       min_per_match                       logval
                   0                            0
```

```
Linear Regression

100 samples
  5 predictor

No pre-processing
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 81, 81, 80, 79, 79
Resampling results:

  RMSE       Rsquared   MAE
  0.4593669  0.7981805  0.3611082

Tuning parameter 'intercept' was held constant at a value of TRUE
```

Appendix D: References

Background

1. DiBlasi, L. (2022, April). Footballer Valuations: Valuing World-Class Football Players Against Transfer Fees [Review of Footballer Valuations: Valuing World-Class Football Players Against Transfer Fees]. Bryant University. https://digitalcommons.bryant.edu/cgi/viewcontent.cgi?article=1042&context=honors_economics
2. Metelski, A. (2021, April 30). Factors affecting the value of football players in the transfer market. ResearchGate. https://www.researchgate.net/publication/351335017_Factors_affecting
3. Poli, R., Ravenel, L., & Besson, R. (2020, March). Scientific evaluation of the transfer value of football players. Monthly Report 53. https://football-observatory.com/IMG/sites/mr/mr53/en/âĂŃ
4. Zhou, W. (2019). Which Factors Decide the Market Value of Soccer Players: An Empirical Evidence from European League. Clausius Press. https://www.clausiuspress.com/conferences/LNEMSS/EMSD%202019/19EMSD057.pdfâĂŃ

Data

1. Players CSV / Football Stats Database to CSV | FootyStats. (2019). Footystats.org. https://footystats.org/download-stats-csvâĂŃ
2. Premier League 18/19. (2018). Transfermarkt.us. https://www.transfermarkt.us/premier-league/startseite/wettbewerb/GB1/saison_id/2018âĂŃ

Supplemental Code and Analysis Help

1. Cheng, M. (2023, February 15). Simple examples of cross-validation. RPubs. https://rpubs.com/muxicheng/1004550