Contributed Factors in Predicting Market Values of Loaned Out Players of English Premier League Clubs

Muhammad Daffa Arviano Putra[1], Deshinta Arrova Dewi[2],
Wahyuningdiah Trisari Putri[3], Retno Hendrowati[4], Tri Basuki Kurniawan[5]
Department of Informatics-Faculty of Engineering Science, Paramadina University, Jakarta, Indonesia[1, 3, 4]
Faculty of Data Science and Information Technology-INTI International University, Nilai, Malaysia[1, 2]
Postgraduate Program of Information Technology-Bina Darma University, Palembang, Indonesia[5]

***Abstract*—The top tier of the English football league division is occupied by the English Premier League (EPL). It has become a global phenomenon with exhilarating skills and has been one of the most-watched professional football leagues on the planet. The possibility of a player temporarily playing for a club other than the one to whom they are now contracted is known as a "loan player" in the English Premier League (EPL) hence, each player has a market value. Market value is an estimate of how much a player costs when a club wants to buy his contract from another club. The purpose of this study is to determine the factors that influence a player's market value at the conclusion of a loan period. With the Transfermarkt player transfer record dataset for the years 2004 through 2020, we use linear regression analysis. Our study found that a football player's market worth at the end of a loan period is influenced by several aspects, including market value at the beginning, goals, appearances, and total loan.**

***Keywords*—Data analytics; predicting market value; English Premier League; loaned out players; consumption; resource use**

I. Introduction

Football is one of the most popular team sports worldwide [1]. According to research conducted by Nielsen Sports, more than 40% of people aged 16 or older in countries with high populations and large markets said they are "interested" or "very interested" in following football [2]. The most prominent and well-known football league in the world is the English Premier League (EPL). The EPL has a lot of fans all over the world. According to the official website of the English Premier League, the cumulative global audience of EPL for season 2018/2019 was over 3 billion [3]. Due to the huge fans and excitement of the EPL, it has succeeded in attracting the interest of investors. Matchday revenue, broadcast deals, and commercial activity are some of the ways a club can generate a lot of profit [4].

To maximize revenue, the club must be popular among the viewers and have winning matches. Therefore, every owner strives to strengthen their club squad to be competitive and have a successful season. One way to strengthen a club is to buy good and talented players. Some of the examples we saw recently when Manchester City bought Jack Grealish from Aston Villa with a figure of €117.50m, Chelsea bought Romelu Lukaku from Inter Milan for €115.00m, and Manchester United bought Jadon Sancho for €85.00m from the German club, Borussia Dortmund [5].

Each player has a market value. Market value is an estimate of how much a player costs when a club wants to buy his contract from another club [1]. The price does not apply if a club only loans a player. If a club wants to loan a player, they are most likely only required to pay the player's salary for the duration of the loan. However, the player will still carry a market value that can go up or down when playing at the loan club. During the loan period, the player might perform extraordinarily and make many appearances and therefore, his market value will increase and vice versa. In every transfer window, every club in the EPL is likely to loan out their players to make room for their squad or give players a chance to build a reputation at another club. For this study, we use the market value published by the Transfermakt website. The market values provided by the website are economically relevant and are viewed as having a fine reputation in the sports industry [6]. A study by Peeters revealed that Transfermarkt crowd valuation is referenced privately by club officials during player contract negotiations because it is more accurate than other valuations such as FIFA ranking and the ELO rating [7].

The academic community has found the football transfer market to be an engaging subject [8]. Additionally, we believe that additional investigation into the market value of football players would be an intriguing topic to pursue. Fortunately for us, the information required to do so is readily accessible through websites devoted to the sport of football. The general contributing aspects to a player's market value are not often covered in academic publications, nevertheless. Since the market value at the conclusion of the loaned time in EPL is greater than average, we are looking for contributing causes for that higher market value in this study.

A. Data Availability Statement

The data collected for the model is from Transfermarkt, a German Website that provides football information and data, such as scores, league tables, club squads, and many more using web scraping. Football-related research has used this website as its source data. The website incorporates crowdsourcing to estimate a player's market value in several professional football leagues. This means that every person can join the community and discuss the market value of any football player.

II. Previous Studies

There were studies on the subject of prediction of the market value of a football player that aligned with this research. Such as one from Singh and Lamba that suggested consistency, popularity, crowd estimation, and performance parameters enhance the prediction accuracy of the market value of football players [9]. While Felipe et. al. stated in their paper that the playing position (attacking midfielders) and age of the player (born in the first quarter of the year) are the most economically valued in terms of current value and maximal value [10]. Müller, Simons, and Weinmann stated that there are three categories of indicators of the market value of a player. There are the player characteristics (age, height, position, footedness, nationality), player performance (playing time, goals, assists, passing, dribbling, dueling, fouls, and cards), and player popularity (news, internet links) [1]. Further study on the popularity component, Frenger et. al. suggested that social media activities significantly influenced a football player's market value on the site [11].

A study on the player performance is also done by Richau et. al. which emphasized the actual performance of a football player to determine their market value, the paper stated that the actual performance is measured through individual player's age, minutes played, offense, defense, and team and analyzed using boosted regression trees [12]. They found that individual player performance indicator does not have the highest influence on the market value; instead, they found that team dimension average rank influences market value. Along this line, Metelski tried to find factors affecting the value of football players in the transfer market for the Polish Football League using descriptive statistics and several statistical tests. The writer uses the position on the pitch, age at transfer, year of transfer, the destination country, and selling club as the indicators. They found that the age of the player is a significant factor in the football players' values [13]. Moreover, Behravan and Razavi used the FIFA 20 dataset of various performance ratings of 18,278 players. Their novelty is the use of an automatic clustering algorithm in the first phase which they called APSO-clustering, and further training of a hybrid regression method called PSO-SVR for each cluster [14] that can estimate the players' value with an accuracy of 74%.

III. Method

Fig. 1 illustrates the methods we use for this study. These steps will be more specifically described afterward.

B. Data Collection

The data collected for the model is from Transfermarkt, a German Website that provides many footballs information and data, such as scores, league tables, club squads, and many more using web scraping. Football-related research has used this website as their source data, such as in [14] [15][16][6][17]. The website incorporates crowdsourcing to estimate a player's market value in several professional football leagues. This means that every person can join the community and discuss the market value of any football player. Everyone can suggest a market value for a player with good arguments and reasons to justify the player's estimation [6]. However, not everyone's opinion has the same value. The Transfermarkt website has data on past loan players' performance in the EPL from the season 2004/2005 to 2021/2022. However, as the EPL 2021/2022 season is still ongoing at the time of the writing of this paper, we will limit the data to season 2020/2021.

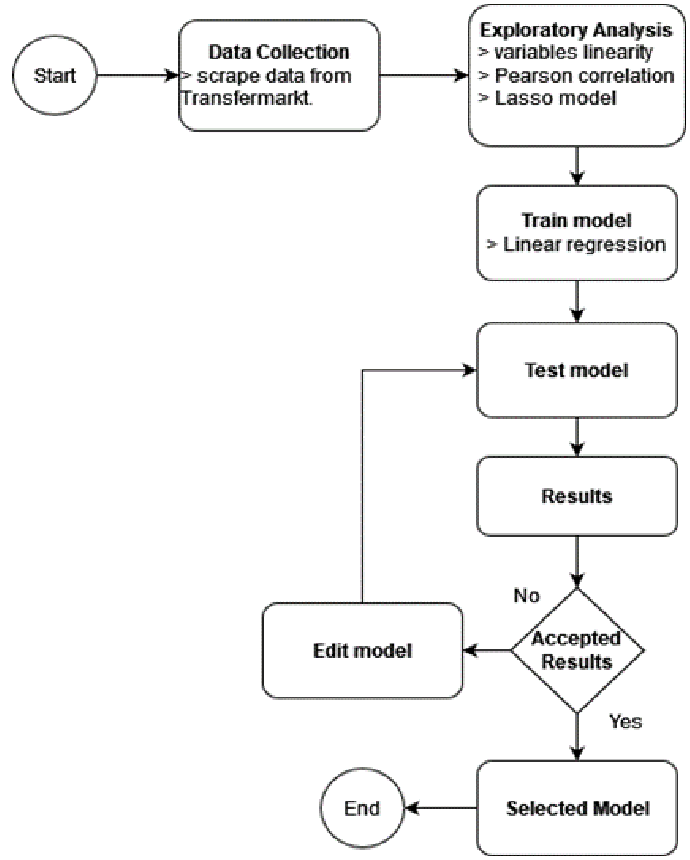


Fig. 1. Flowcharts of an applied method for the research.

There are 10 variables in the data as captured in Table I. In this study, we want to predict the last variable, which is the total market value of the loaned-out players from a club in a season. Therefore, the first nine variables are labeled as the explanatory variables and the last variable is our target or dependent variable.

C. Exploratory Analysis

In this section, we explore the data that we have. Our data has 339 rows and 10 columns, which means we have a total record of 339 data of all clubs who loaned out their players from season 2004/2005 to 2020/2021. Every season consists of 20 clubs and because the EPL uses a promotion and relegation system, there can be more than 20 unique clubs in the data. The first thing we have done is to list and count every unique club in the data. The outcome is that we have 40 unique clubs that will be trained in the Linear Regression model. We start exploring our data with the perspective of the relationship between variables, the data distribution, the correlation among variables, Least Absolute Shrinkage and Selection Operator (LASSO) in identifying features that may be the contributing factors to predicting market values of the loaned players.

TABLE I. SUMMARY OF THE AFM INFORMATION OF CDS QDS

Variable Name	Description
name	The club's name
year	The year of loaned out players data of a club (ranging from 2004 to 2020)
total_loan	The total loaned-out players from the club in the respective year
average_loan (in years)	The average number of loan periods of all loaned out players from the club in the respective year
appearances	The total number of appearances from all loaned-out players from the club in the respective year
starting_formation	The total number of appearances in the starting formation from all loaned-out players from the club in the respective year
goals	The total number of goals from all loaned-out players from the club in the respective year
average_minutes_played	The average minutes played by all loaned-out players from the club in the respective year (the maximum is 90 as a football match is played for 90 minutes)
market_value_at_start (in M €)	The total market values at the start of the loan period from all loaned-out players from the club in the respective year
market_value_at_end (in M €)	The total market values at the end of the loan period from all loaned-out players from the club in the respective year

D. Checking the Relationship between Variables using Scatter Plots

We start by checking the bivariate relationships between eight variables of our data. The total loan personnel, average loan time in years, number of appearances made by the players, starting formation of the players, goals made, average minutes played, market value at the start, and the last variable is the market value at the end. As we can see from the scatter plot in Fig. 2, there are three types of relationships shown, discrete relationships, random, and linear.

The discrete plot was obtained from the total loaned variables as there were only two values in these variables as the players were loaned only for 1 year or 2 years. The random relationship is shown by the total loaned players' variable against average minutes played, market value at the start, and market value at the end. The random relationship we see with average minutes played against all other variables. While the rest of the bivariate combinations showed some kind of linear relationship.

In this paper, we focused on the relationship between the market value at the end with the rest of the variables, so we found that the market value at the end has a strong linear relationship with the market value at the start. We explore the relationship more in the sections below.

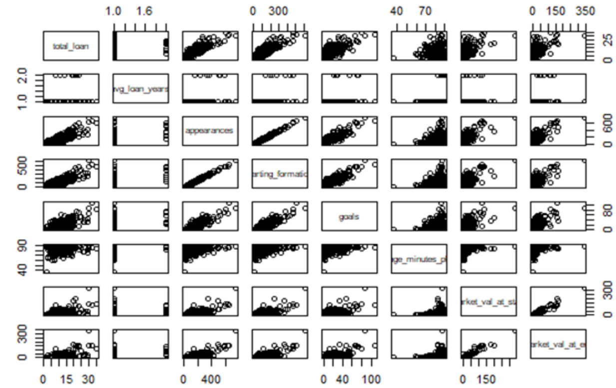


Fig. 2. Bivariate relationship of variables data.

E. Checking the Univariate Distribution using Histogram

To see if the observed data represent a random sample from the population; we use the histogram to check the distribution. Fig. 3 below shows the distribution of seven variables that we consider from the data; we did not include the name and year variables because they are categorical. The initial histogram showed left and right skewed data, so we do a log transformation on the variables to stabilize the variance, such as seen in [18] and [19]. Ensuring the log transformation, the average minutes played variable is still left skewed, while the average loan years is discrete.

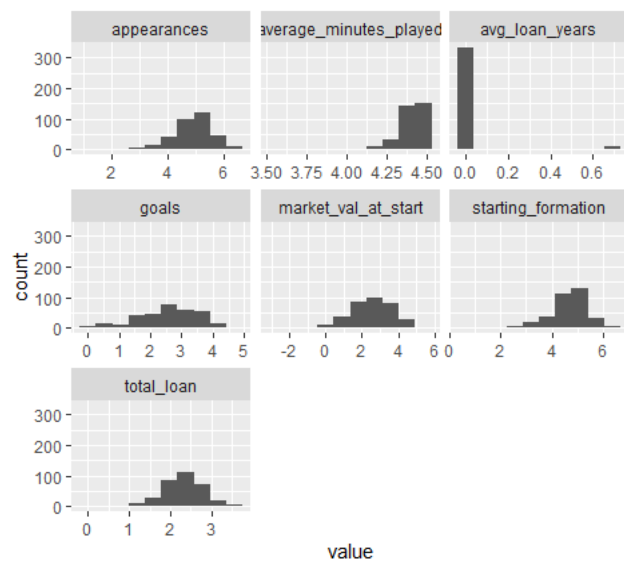


Fig. 3. Log transformed data distribution of market value predictors.

F. Pearson Correlation

The next thing we do in our exploratory step is to see the Pearson correlation to determine the precise extent or degree of any connection between any two variables, indicating its presence or absence.

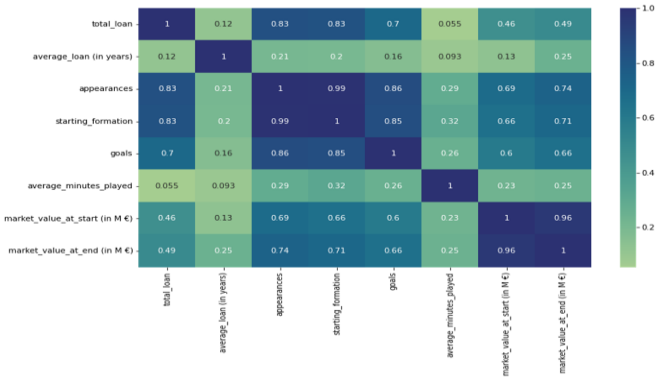


Fig. 4. Pearson correlation among eight variables.

Fig. 4 shows that three variables show a strong correlation to the market value at the end. From this figure, we can first assume that the variables such as market value at the start, appearances, and starting formation may be the contributing factor for our model with the value of 0.96, 0.74, and 0.71 respectively.

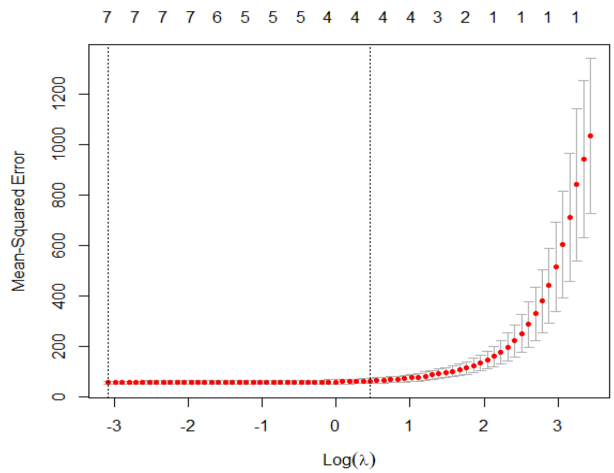
G. Selecting Features using Lasso Model

We investigate the idea of simplifying the model in this part. We want to simplify the model since we're attempting to identify the factors that can accurately forecast a player's market value. Isolating variables is for better prediction can be found in [20], [21], and [22]. We use the Lasso model to reduce the variability of the estimates by shrinking some of the coefficients exactly to zero for an easily interpretable [23]. In this regard, Fig. 5 is considered.

In Fig. 5, we use the λ min option, which refers to the value of λ when the cross-validation has the lowest error. The results showed that the lambda value is 0.04615499. After Lasso, there are still seven predictors. Total loan and average minutes played variable has a negative correlation to market value at the end, and the market value at the start variable has the highest correlation amongst other variables to the market value at the end variable.

H. Train and Test

We fitted our linear regression model using our dataset. We employ 3, 5, and 10-fold cross-validation for the experiment. For the base model (Model 1), we go with the nine predictors from our dataset such as name, year, total loan, average loan, appearances, starting formation, goals, average minutes played, and market value at the start, and then we proceed with the second model (Model 2) with seven predictors, removing the name and year variables. The third model (Model 3) is acquired by removing more variables such as average minutes played and average loan years. We use the metrics MAE, MSE, RMSE, and R2 to measure how our model performs. Table II shows the outcome of our training and testing procedure. We sought lower MAE, MSE, and RMSE values for the training model because they indicate that our model is more accurate. Higher R2 was our goal because it indicates how well the linear regression model captures the variability in the data. From Table II we can conclude that Model 1, Model 2, and Model 3 performed well with their respective numbers of predictors. Model 3 has the highest accuracy, followed by Model 2 and Model 1 in that order. Model 3 has the fewest variables compared to the other models, which indicates a less sophisticated model, which is most likely why this happened.




```
8 x 1 sparse Matrix of class "dgCMatrix"
                                   s0
(Intercept)           -22.85520355
total_loan             -0.28072857
avg_loan_years         26.52136240
appearances             0.01866043
starting_formation      0.01392443
goals                   0.12721553
average_minutes_played -0.06779750
market_val_at_start     0.88581147
```

Fig. 5. Lasso model's result.

TABLE II. MODEL'S TRAIN AND TEST RESULT

c	Model = 1			Model = 2			Model = 3		
	K = 3	K = 5	K = 10	K = 3	K = 5	K = 10	K = 3	K = 5	K = 10
MAE (Negated)	- 6.2674 9912	- 3.4813 1114	- 4.44286 823	- 5.2452 6151	- 3.3743 0405	- 3.7785 9705	- 0.2231 9212	- 3.3743 0405	- 3.7785 9705
	- 4.9252 4257	- 6.4046 8388	- 2.64370 08	- 4.9276 4655	- 5.9135 472	- 2.5008 1717	- 0.2076 8207	- 5.9135 472	- 2.5008 1717
	- 3.9666 9784	- 6.3428 0056	- 10.1452 6452	- 3.9004 8279	- 6.1742 0242	- 9.4957 1448	- 0.1955 6174	- 6.1742 0242	- 9.4957 1448
		- 3.7498 3279	- 2.42966 98		- 3.4476 2154	- 2.5524 5853		- 3.4476 2154	- 2.5524 5853
		- 4.2139 4538	- 5.39165 79		- 4.1220 4388	- 5.1203 4287		- 4.1220 4388	- 5.1203 4287
			- 7.43961 965			- 7.0215 2678			- 7.0215 2678
			- 3.10051			- 3.0109			-

			004 -4.3130265 -4.61836063 -3.87216171			2936 -3.6762294 -4.29507904 -3.93445955			3.01092936 -3.6762294 -4.29507904 -3.93445955
MSE (Negated)	-154.7218799 -62.06658612 -33.27916393	-22.92409381 -188.59043175 -96.45110892 -23.38062198 -41.2124202	-28.00937513 -15.52227051 -327.00278803 -9.75251004 -53.57044859 -150.95979864 -15.88412907 -29.84904304 -45.07887223 -38.79109755	-114.3565745 -61.14854143 -33.52075828	-20.76053851 -153.389980311 - 92.10255218 -21.8087315 -40.31789501	-24.43446739 -12.97297629 -276.861121156 -10.46169465 -51.01546795 -126.115700599 -15.48585609 -25.65816041 -42.11266768 -38.42064157	-0.17374381 -0.11261894 -0.06139012	-20.76053851 -153.389980311 -92.10255218 -21.8087315 -40.31789501	-24.43446739 -12.97297629 -276.86112156 -10.46169465 -51.01546795 -126.11570059 -15.48585609 -25.65816041 -42.11266768 -38.42064157
RMSE (Negated)	-12.43872501 -7.87823496 -5.76880958	-4.78791122 -13.73282315 -9.82095255 -4.83535128 -6.41969005	-5.29238841 -3.93983128 -18.08321841 -3.1229009 -7.3191836 -12.28656985 -3.98548982 -5.46342777 -6.71408015 -6.22824996	-10.69376335 -7.81975328 -5.78971142	-4.55637339 -12.38507086 -9.59700746 -4.66998196 -6.34963739	-4.94312324 -3.60180181 -16.63914426 -3.2344543 -7.14251132 -11.23012469 -3.93520725 -5.06538848 -6.48942738 -6.19843864	-0.41682587 -0.33558746 -0.2477703	-4.55637339 -12.38507086 -9.59700746 -4.66998196 -6.34963739	-4.94312324 -3.60180181 -16.63914426 -3.2344543 -7.14251132 -11.23012469 -3.93520725 -5.06538848 -6.48942738 -6.19843864
R^2	0.93104303 0.90090885 0.8371551	0.93765187 0.94371147 0.86964737 0.82192481 0.82349607	0.93943311 0.91363223 0.94060919 0.97748276 0.87517002 0.85514786 0.77168829 0.82540351 0.79601323 0.82879776	0.94903317, 0.90237453, 0.83597291	0.94353623 0.95421774 0.87552439 0.83389689 0.82732713	0.94716342 0.9278168 0.94971601 0.97584535 0.88112364 0.87898679 0.77741289 0.84991731 0.80943563 0.83043275	0.90670247 0.89074606 0.9162714	0.94353623 0.95421774 0.87552439 0.83389689 0.82732713	0.94716342 0.9278168 0.94971601 0.97584535 0.88112364 0.87898679 0.77741289 0.84991731 0.80943563 0.83043275

IV. Results and Discussion

The exploratory analysis revealed that there was random and linearity in the bivariate connection between the variables, indicating that the data is suitable for the linear regression model. The log-transformed variables showed that only average minutes played were still right-skewed suggesting that this metric may not be a reliable indicator of final market worth. The discrete variable average loan years are another one that may not be a good predictor.

The Pearson correlation further revealed that three variables strongly correlated with the final market value. They are the market value at the start, appearances, and starting formation. We can assume that these three variables are good predictors. The LASSO model confirms that 7 variables are fairly good for predicting market value at the end. However, we still suspect that the right skewed and discrete distribution is not a good predictor, hence we removed the two predictors (average minutes played and average loan years) and came up with model 3.

Based on Table II, model 1 is the base model with nine predictors, model 2 with seven predictors, and model 3 with five predictors. The train and test for all three models showed that model 1 has the least R2 values, at an average of 88.9%, 87.9%, and 87.2% for respective 3, 5, and 10 folds. The percentage showed how well the dependent variable, market value at the end can be accounted for by the nine predictors. Model 2 came second with the average R2 values of 89.6%, 88.7%, and 88.3% accounted for by seven predictors, and model 3 has the highest average R2 values of 90.5%, 90.3%, and 90.1% accounted for by five predictors.

In terms of accuracy, the mean RMSE for model 1 for 3, 5, and 10 folds, respectively are 8.69525, 7.91934, and 7.24353. For Model 2, the mean RMSE are 8.10107, 7.51161, and 6.84796, respectively. For Model 3, the mean RMSE are 0.33339, 0.32948, 0.31620, respectively. Model 3 accuracy is higher than the two previous models. Therefore, we can conclude that removing the two variables average minutes played, and average loan years, is the right decision. With model 3 we come up with this linear equation:

$$market\_val\_at\_end = (-2.282 - 0.04653 * total\_loan) + (0.5237 * appearances) - (0.00007325 * starting\_formation) + (0.1056 * goals) + (0.8716\, market\_val\_at\_start) \quad (1)$$

With the use of the aforementioned equation, we can observe that while every variable affects the market value at the end, the market value at the beginning, objectives, appearances and total loan all has positive correlations and significant contributions. This differs slightly from our initial hypotheses based on the Pearson correlation, which was initial market value, early appearances, and initial formation.

The majority of the contributing factors to the loan player in the ELP have generally been identified by our investigation. With this investigation and its outcome, we have discovered a previously unknown association. We have identified which variables are connected to or have the strongest relationships with, and we may be able to identify patterns within the dataset as a result of this understanding.

V. Conclusion

In this study, we identified the elements that contributed to a football player's market worth in the English Premier League at the end of the loan period. Market value at launch, goals, appearances and total loan is among them. The market worth of a football player after a loan has been made is something we can forecast using exploratory research and a linear regression model. To discover the best predictor, we tested three distinct models. Our study revealed that the predictors differed slightly from what we had initially thought. Although linear regression is simple to comprehend and explain, we believe the model is adequate for use in this investigation. In future experiments, we hope to incorporate more data into our model, as the current data is very limited. We can also implement other models to better understand the contributing factors of a football player's market value.

References

[1] O. Müller, A. Simons and M. Weinmann, "Beyond crowd judgments: Data-driven estimation of market value in association football," *European Journal of Operational Research,* vol. 263, no. 2, pp. 611-624, 2017.

[2] "World Football Report 2018," Nielsen Sports, 2018. [Online]. Available at: https://www.nielsen.com/wp-content/uploads/sites/3/2019/04/world-football-report-2018.pdf.

[3] "Premiere League Global Audience on The Rise," Premier League, 2019. [Online]. Available: https://www.premierleague.com/news/1280062.

[4] T. Dima, "The Business Model of European Football Club Competitions," Procedia Economics and Finance, vol. 23, no. Oct. 2014, pp. 1245-1252, 2015.

[5] "Premiere League - Transfer records," Transfermarkt, 2021. [Online]. Available: https://www.transfermarkt.com/premier-league/transferrekorde/wettbewerb/GB1/plus//galerie/0?saison_id=2021 &land_id=alle&ausrichtung=&spielerposition_id=alle&altersklasse=&leihe=&w_s=s&zuab=zu.

[6] S. Herm, H.-M. Callsen-Bracker and H. Kreis, "When the crowd evaluates soccer players' market values: Accuracy and evaluation attributes of an online community," Sport Management Review, vol. 17, no. 4, pp. 484-492, 2013.

[7] T. Peeters, "Testing the Wisdom of Crowds in the field: Transfermarkt valuations and international soccer result," International Journal Forecasting, vol. 34, no. 1, pp. 17-29, 2018.

[8] D. Matesanz, F. Holzmayer, B. Torgler, S. L. Schmidt and G. J. Ortega, "Transfer market activities and sportive performance in European first football leagues: A dynamic network approach," PLOS ONE, vol. 13, no. 12, pp. 1-16, 2018.

[9] P. Singh and P. S. Lamba, "Influence of crowdsourcing, popularity and previous year statistics in market value estimation of football players," Journal of Discrete Mathematical Sciences and Cryptography, vol. 22, no. 2, pp. 113-126, 2019.

[10] J. L. Felipe, A. Fernandez-Luna, P. Burillo, L. E. de la Riva, J. Sanchez-Sanchez and J. Garcia-Unanue, "Money Talks: Team Variables and player Positions that Most Influence the Market Value of Professional Male Footballers in Europe," Sports Policy and Finance, vol. 12, no. 9, pp. 10-17, 2020.

[11] M. Frenger, F. Follert, L. Richau and E. Emrich, "Follow me... on the relationship between social media activities and market values in the German Bundesliga," Saarbrücken, 2019. [Online]. Available: http:www.soziooekonomie.org.

[12] L. Richau, F. Follert, M. Frener and E. Emrich, "Performance indicators in foorball: The importance of actual performance for the market value of football players," SCIAMUS Sport and Manag, vol. 4, pp. 41-61, 2019.

[13] A. Metelski, "Factors affecting the value of football players in the transfer market," Journal of Physical Education and Sport, vol. 21, no. 2, pp. 1150-115, 2021.

[14] I. Behravan and S. M. Razavi, "A novel machine learning method for estimating football players' value in the transfer market," Soft Computing, vol. 25, no. 3, pp. 2499-2511, 2021.

[15] H. Adiwiyana, H. I. Adiwiyana and Harywaman, "Factors that Determine the Market Value of Professional Football Players in Indonesia," J. Din. Akunt, vol. 13, no. 1, pp. 51-61, 2021.

[16] R. Stanojevic and L. Gyarmati, "Towards Data-Driven Football Player Assessment," IEE Int. Conf. Data Min. Work. ICDMW, vol. 0, pp. 167 172, 2016.

[17] M. He, R. Cachucho and A. Knobbe, "Football Player's Performance and Market Value," in Proc 2nd Work. Sport. Anal. Eur. Conf. Mach. Learn. Princ. Pract. Knowl. Discov. Databases (ECML PKDD), 2015.

[18] D. Curran-Everett, "Explorations in statistics: the log transformation," Adv. Physiol. Educ., vol. 42, no. 2, pp. 343-347, 2018.

[19] H. Son, C. Hyun, D. Phan and H. J. Hwang, "Data analytic approach for bankruptcy prediction," Expert Systems with Applications, vol. 138, 2019.

[20] Z. Yan and Y. Yao, "Variable selection method for fault isolation using least absolute shrinkage and selection operator (LASSO)," Chemom. Intell. Lab. Syst., vol. 146, pp. 136-146, 2015.

[21] S. Tian, Y. Yu and H. Guo, "Variable selection and corporate bankruptcy forecasts," Journal of Banking & Finance, vol. 52, no. December, pp. 89-100, 2015.

[22] P. Ghosh, S. Azam, M. Jonkman and A. Karim, “Efficient Prediction of Cardiovascular Disease Using Machine Learning Algorithms with Relief and LASSO Feature Selction Techniques,” IEEE Access, vol. 9, pp. 19304-19326, 2021.

[23] M. R and R. R, “LASSO: A Feature Selection Technique in Predictive Modeling for Machine Learning,” 2016 IEEE International Conference on Advances in Computer Application (ICACA), pp. 18-20, 2016.