2022 International Conference on Trends in Quantum Computing and Emerging Business Technologies (TQCEBT) | 978-1-6654-5361-5/22/$31.00 ©2022 IEEE | DOI: 10.1109/TQCEBT54229.2022.10041664

A Non-Linear Approach to Predict the Salary of NBA Athletes using Machine Learning Technique

Atishay Jain
Research Scholar
CHRIST (Demeed to be University),
atishay.jain@msds.christuniversity.in

Pancinovia Neelu M
Research Assistant
CHRIST (Demeed to be University)
pancinovia.neelu@christuniversity.in

Shreyans Jain
Research Scholar
CHRIST (Demeed to be University),
shreyans.jain@msds.christuniversity.in

Jossy P George
Professor
CHRIST (Demeed to be University)
frjossy@christuniversity.in.

***Abstract*— Every sportsman traded/drafted receives monetary compensation in accordance with their contract. In this study, we propose a nonlinear approach based on performance and other aspects to determine the salary of a basketball player. We estimate the salary based on four regressive models. Whilst predicting we also figure out the important features impacting the salary. Comparatively speaking, random forest outperformed other algorithms. Furthermore, we consider that our findings might benefit discussions between basketball teams and players. This model can also help set a benchmark for salary expectations by the players in accordance.**

***Keywords*— Non-Linear Model, Regression machine learning, Salary prediction, Sports analytics, XGBoost, Random Forest**

I. Introduction

This National Basketball Association (NBA) is a professional sporting league in North America and is the most renowned basketball league in the world. One of the top professional sports leagues in the United States and Canada, the league has 30 teams. The teams are divided into two conferences – Eastern and Western. The teams from each conference (East and West) are ranked based on number of games won or lost. Based on their ranking the top six teams qualify for the playoffs. Teams with ranks 7-10 enter a Play-In Tournament from each conference. And the top two teams advance to the Playoffs of their respective conference. In the end, teams from each conference play against one another to determine who is the best team based on wins. After which a series of seven matches are held amongst the finalists of Western conference and Eastern conference. This decides the ultimate champion team of the particular year or season.

In recent times, NBA's business worth is on a stride with millions of viewers and commercials. Given the athlete's popularity and demand, management's perspective is centered on the decision to select the player. Although modern technology and sports analytics provide great support to pick players. Choosing them has a cost.

The cost can be thought of as salary plus other benefits. One core component based on which a player is chosen by a team is their performance. Other attributes like Age, previous salary, position, fantasy points, etc. can also be utilized to pick a player for the team. Machine learning can be deployed to comprehend patterns that enable analysis and prediction. Such results can be exploited by the Team manager to negotiate monetary terms with the player.

There exists widespread research by various researchers to predict the salary of an athlete based on several aspects. However, the crucial fact is to check if linearity is viable before predicting the salary. Most of the models attempt to find a relationship based on attributes and salary forgoing linearity or nonlinearity present between the variables. The first model that is often run is multiple linear regression, which does not appear to be a viable model to forecast since there may be a non-linear relationship between the qualities and pay. we focus on nonlinear regressive models to predict pay in this study. according to our findings, the random forest regressor is a superior predictor.

The rest of the paper is organized as follows. The literature review is narrated in the second section. Third section explains the Methodology. Followed by data-set explanation and Machine Learning Algorithms in fourth and fifth section respectively. Evaluation Metrics & Results are explained in sixth section. Lastly the conclusion is stated in seventh section.

II. Literature review

Mustafa A.Al-Asadi and Sakir Tasdemir [1] demonstrated how machine learning algorithms may be used to forecast football player worth. Transfer fees that may be paid for a football player on the football market provide a decent indication of a player's market value . Traditionally the market has been valued by football experts. However, the judgments are not accurate. The study makes use of FIFA 20 video game data obtained from sofifa.com. . Four regression models—linear regression, multiple linear regression, decision trees, and random forests—that were tested on the entire collection of variables were used to estimate market worth for the players. Eventually, they believe that our findings can considerably assist conversations between football teams and a player's agency. They conclude in their study that these models can be utilized as a starting point to quantitatively determine a player's market value.

Nuoya Li [2] establishes a link between on-court performance, individual traits, and pay in addition to conducting two regression analyses were performed to investigate the factors that influence compensation and overpayment in the first year of a new contract. In the first regression, he observed that a player's wage is impacted by his or her performance not just during the contract year but also the previous year. Through the experiment, he found that the height of a player is an important factor than the weight of a player while predicting the salary. In his second regression, he discovered that providing a sizable contract always resulted in an overpayment in the year, contract was signed. Additionally, he discovered that giving an all-star player a contract could assist the organization,the year after they signed it, but giving an older player a contract would not be the best move.

Fadi Thabtah, Li Zhang et.al., [3] - The first step is to identify the crucial statistics that include the most details about the players' pay. Secondly, we predict the NBA player salaries using the selected statistics. The Least Absolute Shrinkage and Selection Operator (LASSO) was used to select the essential data, and the Random Forest (RF) approach was employed to estimate wages. They demonstrated the need to use non-linear models and algorithms since NBA player statistics and compensation have a clearly non-linear relationship. Additionally, they demonstrated how to quantify a model's predictability as well as the proper method for examining the link between a response and a large number of predictor variables. The RandomForest non-linear algorithm was used to forecast player salaries after the researchers used LASSO variable selection to identify the significant elements (statistics) that are most frequently linked to NBA player's salary salaries. The accuracy of forecasts made about NBA player wages based on the players' on-court performance is acceptable, but not as good as one might like. Researchers argue that important characteristics included in the study, such as popularity, the quantity of spectacle offered, and so on, might significantly improve the accuracy of pay projections.

III. Methodology

The method employed in this work is supervised machine learning. Another pointer to keep in mind is that. in Machine Learning we have the actual values of the variable we wish to predict. In such a learning technique the algorithm learns from a part of the dataset to come up with a model (also known as training the model).

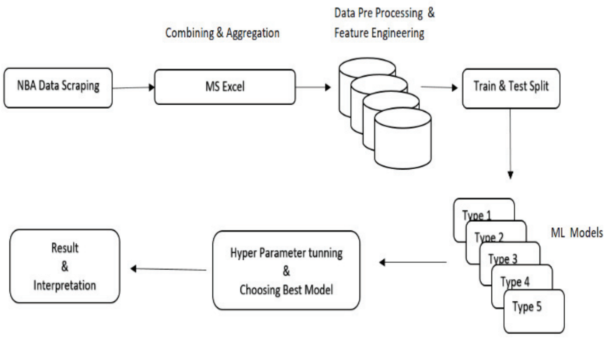

Fig. 1.   Proposed Framework for NBA Salary Prediction

The model is then put to test using the other portion of the dataset that wasn't used to build the model. Finally, the predicted values are compared with actual values to quantify the model's accuracy for prediction. The methodology used is shown in Fig 1. The below are the steps followed for optimal model.

A. *Scrutiny of variables / attributes to understand the factors that impact player's salary while predicting:*

In this step, we went through studies that predict a player's salary. In addition, we examined studies relating to factors affecting salary. Performance, age, Position, and fantasy points were various categories that impact salary based on the Literature study. Performance can be measured based on various factors like minutes played, 3 pointers made, shots attempted, blocks, assist, fouls, steals, etc. Various performance-based factors can be seen in NBA Advanced Stats.

B. *Data Scrapping & Aggregation*

This step mainly deals with web scrapping using python. Scrapping is an automatic process to retrieve large amount of data from websites. Selenium, Beautiful Soup and Pandas packages were used to web scrape data from the official website of the NBA. However, the salary of respective players was taken from ESPN's official website.

C. *Pre Processing & Feature Engineering*

Pre-processing is a crucial step for the transformation and preparation of data to an apt form suitable for building the predictive model. It comprises procedures such as data processing, cleansing, and minimization [5]. Cleaning lays the way for more accurate and reliable findings. As a result, the essential steps were done to clean and analyze the data for modeling.

D. *Exploratory Data Analysis*

To study the relativeness and quality of the features with salary (the value to be predicted) Pearson Correlation Coefficient was used. In addition, Feature importance function was used from "scikit-learn" in Python. In general, valuable features tend to be highly correlated with the variable to be predicted, salary in our case. Fig 2 showcases

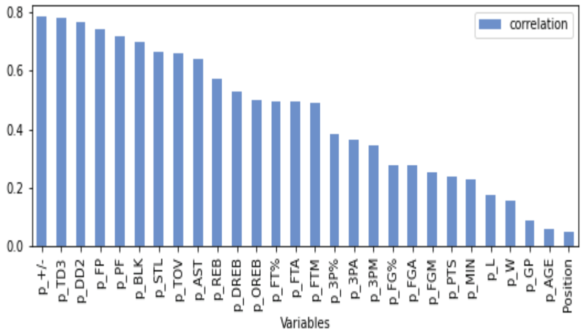

Fig. 2.   Correlations of player's Performance Measures to Salary of Players

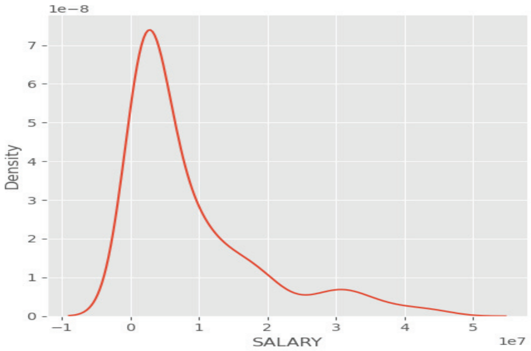

Fig. 3.   Distribution of Salary of Players

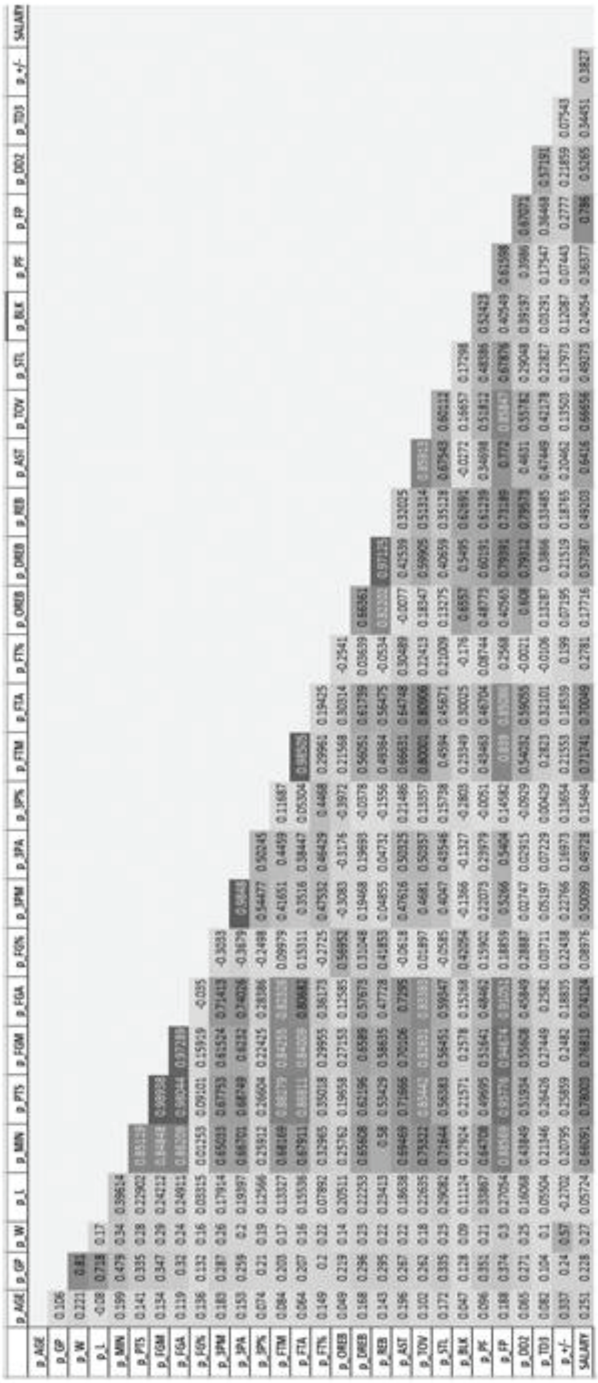

Fig. 4. Correlations between Performance Meausres of Player

the correlations of features to the target variable (Salary). Apart from this, the independent features possess correlations amongst each other. This can be observed in Fig 4. the distribution of our target variable is shown in Fig 3.

E. Data Splitting

After cleaning & pre-processing the data (required features /variables), splitting is performed into two broad sets - training & testing. In common parlance, the split ratio for training the model is between 70% to 80% and for testing is 20% to 30%. For this study, we split the data randomly into 80% and 20% train and test the model respectively.

F. Building various models to predict the salary

We estimated the player's salary using various regression models (multiple linear regression, decision trees, random forests, and XGBoost). These models can be built from the module called scikit-learn in Python.

G. Hyper Parameter tuning

Various measures like Root mean square errors (RMSE), Mean absolute errors (MAE), and the coefficient of determination (R2) was utilized to gauge the accuracy of the models with the testing dataset. However, before gauging the model, hyperparameter tuning was carried out. A parameter used to regulate how the model learns is known as a hyperparameter and was used to regulate the learning process. Tuning them is the process of choosing the optimal hyperparameters with an aim of improving the ability of a model to predict

IV. Dataset description

The dataset used was scrapped from NBA and ESPN's official website as mentioned earlier for the year 2020-21. Stats of previous years were considered in accordance with the salary for next year as shown in Table.1. The reason for this is that a player's estimated salary is determined before the season begins. Moreover, players who did not have physical minutes played in the game of the regular season for the year 2020-21 were not taken into consideration as their performance attributes will not be available.

TABLE I. Notation and Meaning of Features for Modeling

Notation	Meaning
+/-	Plus-Minus
3P%	3 Point Field Goal Percentage
3PA	3 Point Field Goals Attempted
3PM	3 Point Field Goals Made
Age	Age
AST	Assists
BLK	Blocks
DD2	Double Doubles
DREB	Defensive Rebounds
FG%	Field Goal Percentage
FGA	Field Goals Attempted
FGM	Field Goals Made
FP	Fantasy Points
FT%	Free Throw Percentage
FTA	Free Throws Attempted
FTM	Free Throws Made
GP	Games Played
L	Losses
MIN	Minutes Played
OREB	Offensive Rebounds
PF	Personal Fouls
PTS	Points
REB	Rebounds
STL	Steals
TD3	Triple Doubles
TOV	Turnovers
W	Wins

The performance characteristics (independent variables) were scaled to unit variance using a scaler function centered on the mean. This was done to get rid of the range (as each variable has a different range) and units. For example Age ranges from 20-40 whereas assist ranges between 0-11. Second, age is represented in years, while minutes played are recorded in minutes. Moreover, the position is a categorical feature that was converted into numeric using Label Encoding.

The ability of performance metrics and other attributes to predict salary has advanced in past decades [11]. Besides, enormous efforts have been invested to analyze the performance attributes of basketball athletes to enable managers to avoid overpaying the player. NBA datasets exhibited effectiveness in predicting player's salaries and other analysis. We considered player statistics of the year 2020-2021 (the previous year) to predict the salary of 2022 as a player's contract is signed during the start of the season.

V. Machine Learning Algorithms

To predict the player's salary based on performance attributes and other attributes of a basketball player, various Supervised Learning methods were cast. Because the data contains expected output values, supervised machine learning algorithms are used. All of these strategies try to capture and use the relationship between pay prediction and performance and other parameters to forecast. Since the dataset is numeric in nature, Regression techniques are used. Multiple Linear Regression, Decision Tree Regressor, Random Forest Regressor, and XGBoost Regressor are the models deployed. Few of them are frequently used in other studies for describing players and data mining domains [4]. Additionally, algorithms were compared between nonlinear models (such as random forest, and decision trees) and linear models (such as multiple linear regression). Linear regression in statistics is the establishment of a linear technique to forecasting based on the relationship between a dependent and one independent variable (also known as response and predictor variables).

A. *Multiple Linear Regression*

Multiple linear regression is a variant of linear regression that includes more than one independent or explanatory variable. . In simple terms, it is the predictor (independent) and predicted (dependent) variables. The model aims to find the best coefficients (betas) for the variables to predict the outcome as close as possible keeping in mind the residual or error is minimum.

The below equation represents multiple linear regression in notation format.

$$y = b0 + b1 * x1 + b2 * x2 + \varepsilon \qquad \text{-(I)}$$

B. *Decision tree Regressor*

Decision tree Regressor learning is a common predictive modeling technique used in machine learning. In decision tree, a model predicts by learning simple decision guidelines and splits the data into nodes with branches constructed on the basis of the best importance from the data. Decision tree regressor works on numeric data wherein the dependent variable are continuous numeric values. Decision trees are widely used machine learning algorithms due to their lucidity [5].

C. *Random Forest Regressor*

Random Forest Regressor is an ensemble learning technique that works by training multiple decision trees and then outputs the average prediction (regression) of those trees. In general, Random forests shows better results compared to the decision tree. However, features can impact its performance [6] .

D. *XGBoost Regressor*

XGBoost Regressor is also an ensemble learning technique but it focuses on efficient implementation of gradient boosting. Different decision trees are constructed and these trees are added orderly to the ensemble and fit to modify the errors and take corrective measures for the previous models.

VI. Evaluation Metrics & Results

Multiple linear Regression model banks on linearity as an assumption and based on the results, the r squared value showcasing accuracy is merely 57%. The results are as below shown in Fig.5.

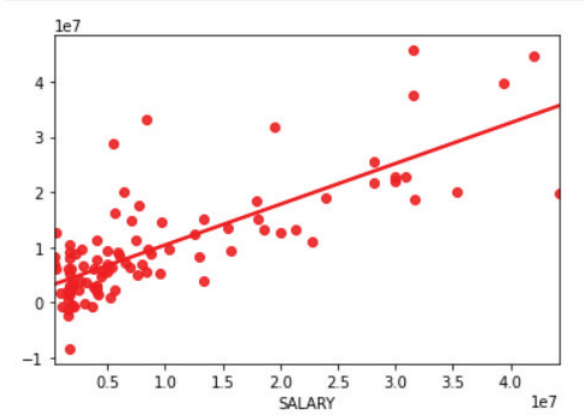

Fig. 5. Actual vs Predicted Salary

Variance Inflation Factors were calculated to check for autocorrelation. Fig.6 shows fantasy points and points have a strong correlation between them. This violates the assumption of Linear Regression. The other models used are Non-Linear and thus show better results.

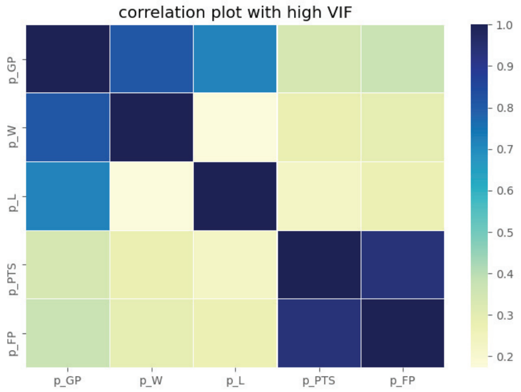

Fig. 6. Correlation within a Player's Performance Metric with High VIF

Decision Tree showed an improvement in the accuracy and the error. The accuracy recorded is 64% upgrading from 57%. A nonlinear link between pay and a few factors is seen in our sample. This explains why the decision tree regressor outperforms linear regression.

On the other XGB Regressor uses ensemble learning with boosting technique. While using XGBoost one tree is built at a time so that the data relating to the respective decision tree is taken into account and missing values are accounted for. This helps in working with gradient algorithms along with the decision tree algorithm which improves results. The R2 value turned out to be higher compared to the Regression tree i.e. 75.85%. The below image shows the relation amongst the predicted and actual values of the salary while using XGB Regressor

Finally, Random Forest leads to the accuracy score of R2 is 88. 18%. The random forest algorithm prevents the model from overfitting by using multiple trees as shown in Table.2. The Root Mean Squared Error (RMSE) is the lowest. Each tree that is produced is based on a separate sample of data. While splitting, a separate sample of characteristics is picked at each node. Then, individual tree models make their predictions. Finally, these forecasts are averaged to get a single outcome. Moreover, while building the random forest model feature importance was checked to get an optimum model as shown in fig.7 and fig.8. Diagrammatic representation is as below.

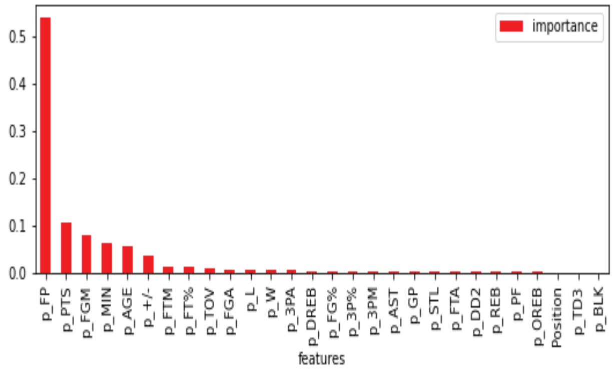

Fig. 7. Importance of Features in Predicting Player's Salary

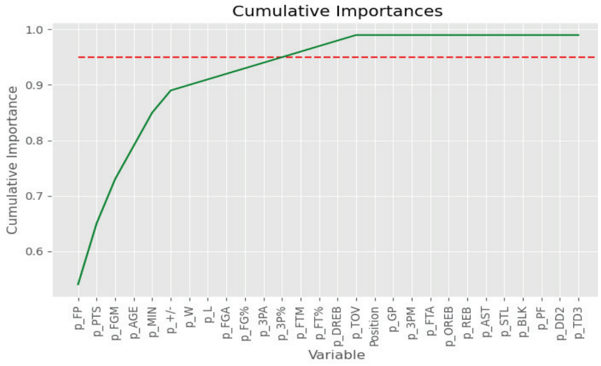

Fig. 8. Cumulative Importance of Features in Predicting Player's Salary

Table1 shows comparison of The R2 and Root Mean Squared Error across the models. The closer the value of R^2 to 1, better the accuracy of the model. Inversely, lower the value of RMSE better the model as it shows smaller error values while predicting the model.

TABLE II. COMPARISON ACROSS MODELS

Model Name	***Split Ratio***	***RMSE***	R^2
Multiple Linear Regression	80:20	69,67,066.5	57.04%
Decision Tree	80:20	57,71,902.35	64.4%
XGB Regressor	80:20	47,54,531.87	75.85%
Random Forest	80:20	33,25,775.78	88. 18%

VII. CONCLUSION

The results of our study specified the viability of a nonlinear model over a linear model while predicting the salary. There was also a breach of the Linear Regression assumption. Thus, the contributions of this work are not limited to pay prediction but also to the methodology utilised in contrast to the conventional strategy used to solve comparable problems (based on prior studies). Our model recorded the best available accuracy rate which depicts that the previous season's performance statistics are a good measure to predict the salary of the current year. The most important variable in predicting the salary was the fantasy points of a player as it showed the feature importance of 54 percent. Future study might focus on estimating wage for the next n years using a neural network if adequate data is available for deep learning. Furthermore, features such as the ratio of games played to losses, time spent at gym, marketing skills, and team playing ability can be utilized to improve salary prediction.

REFERENCES

[1] M. A. A. ASADI and S. TASDEMİR, "Predict the Value of Football Players Using FIFA Video Game Data and Machine Learning Techniques".

[2] N. Li, "The Determinants of the Salary in NBA and the Overpayment in the Year of Signing a New Contract".

[3] F. Thabtah, L. Zhang and N. Abdelhamid, "NBA Game Result Prediction Using Feature Analysis and Machine Learning".

[4] V. Rao and A. Shrivastava, "Team strategizing using a machine learning approach".

[5] X. Wu, V. Kumar, J. R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, P. S. Yu, Z.-H. Zhou, M. Steinbach, D. J. Hand and D. Steinberg, "Top 10 algorithms in data mining".

[6] T. K. Ho, "Random decision forests".

[7] P. Singh and P. S. Lamba, "Influence of crowdsourcing, popularity and previous year statistics in market value estimation of football players".

[8] H. Saikia, D. Bhattacharjee and A. Bhattacharjee, "Performance based market valuation of cricketers in IPL".

[9] N. H. Nguyen , D. T. A. Nguyen, B. Ma and J. Hu, "The application of machine learning and deep learning in sport: predicting NBA players' performance and popularity".

[10] A. S. Markovits and A. I. Green, "FIFA, the video game: a major vehicle for soccer's popularization in the United States".

[11] W. S. Bhaya, "Review of data preprocessing techniques".

[12] M. . A. A. Asadi and S. Tasdemír, "Empirical Comparisons for Combining Balancing and Feature Selection Strategies for Characterizing Football Players Using FIFA Video Game System".

[13] S. H. M. Bracker and H. , "When the crowd evaluates soccer players' market values: Accuracy and evaluation attributes of an online community".