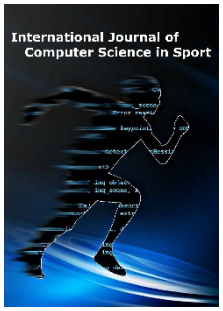
International Journal of Computer Science in Sport

Volume 16, Issue 1, 2017

Journal homepage: http://iacss.org/index.php?id=30

DOI: 10.1515/ijcss-2017-0002



Computational Estimation of Football Player Wages

Yaldo, L., Shamir, L.

Lawrence Technological University

Abstract

The wage of a football player is a function of numerous aspects such as the player's skills, performance in the previous seasons, age, trajectory of improvement, personality, and more. Based on these aspects, salaries of football players are determined through negotiation between the team management and the agents. In this study we propose an objective quantitative method to determine football players' wages based on their skills. The method is based on the application of pattern recognition algorithms to performance (e.g., scoring), behavior (e.g., aggression), and abilities (e.g., acceleration) data of football players. Experimental results using data from 6,082 players show that the Pearson correlation between the predicted and actual salary of the players is ~0.77 ($p < .001$). The proposed method can be used as an assistive technology when negotiating players salaries, as well as for performing quantitative analysis of links between the salary and the performance of football players. The method is based on the performance and skills of the players, but does not take into account aspects that are not related directly to the game such as the popularity of the player among fans, predicted merchandise sales, etc, which are also factors of high impact on the salary, especially in the case of the team lead players and superstars. Analysis of player salaries in eight European football leagues show that the skills that mostly affect the salary are largely consistent across leagues, but some differences exist. Analysis of underpaid and overpaid players shows that overpaid players tend to be stronger, but are inferior in their reactions, vision, acceleration, agility, and balance compared to underpaid football players.

KEYWORDS: FOOTBALL, SOCCER, SPORTS ECONOMY, MACHINE LEARNING

Introduction

Football is by far the world's most popular team sport with ~3.5 billion fans worldwide (Giulianotti, 2012). Due to its popularity, the demand for star football players has been increasing substantially in the past few decades, and football players are being traded by amounts exceeding €100M. These numbers are substantially higher than historical trade figures compared to normal inflation rate. For instance, Diego Armando Maradona, considered the most prominent football player during the 1980s, was traded in his career's prime from F.C. Barcelona to Napoli for a world record fee of $10.48M in 1984 (Frick, 2007), comparable to $24.57M (€22.7M) in 2017 value.

Naturally, the rising transfer rates of football players also affect player's salaries, which have been increasing consistently (Frick, 2007). Determining the compensations for a football player is a complex task in sports economy, affected by a combination of numerous aspects. In the pre-information era that task was performed using mainly qualitative analysis, also because football player statistics and data were not collected consistently, making it more difficult to compare the performance and skills of football players in a quantitative manner (Frick, 2006). However, starting the late 1990s, football data were being collected and published, and the comprehensiveness of these data has been increasing consistently. Public football data currently includes multiple performance figures as well as the salaries of the athletes in all major football leagues (Frick, 2006). In most cases, the superstar football players earn more compared to the other players, also for the income they generate for their clubs through tickets, merchandise, and broadcasting agreements. That effect is magnified by the limited supply of superstar football players, leading to an increased wage through monopsony rents (Garcia-del Barrio and Pujol, 2007). Monopsony rents are predicted in a labor market where multiple employers compete for the services of a limited number of employees (Garcia-del Barrio and Pujol, 2007). In the monopsony football economy, the football clubs compete for the services of a limited number of star athletes, forcing the clubs to increase the wages in order to compete with the other clubs and hire these players.

From a behavioral point of view, the performance of the football players improves when the absolute income of the player increases (Torgler and Schmidt, 2007). On the other hand, it has also been shown that salary inequality has a negative effect on the team, and the performance of a football player declines when the income difference between the players and the rest of the team gets larger (Torgler et al., 2006). Additionally, the salary of the football player has a clear impact on coaching decisions, observed by the fact that players with higher salaries tend to be used by the coaching staff in a fashion that is not proportional to their performance on the field compared to other players who are compensated less generously (Garcia-del Barrio and Pujol, 2009). Salary discrimination is far more dominant than other possible types of discrimination such nationality (Garcia-del Barrio and Pujol, 2009).

While the effort invested by the player has an insignificant effect on the salary (Wicker et al., 2013), it has been proposed that the determination of the football player's salary depends on the players performance in the previous season, games played internationally, and the number of goals scored by the player (Frick, 2011). While these factors clearly affect the football player's wage decision, it is reasonable to assume that other factors such as passing skills, free kick accuracy, speed, and tackle ability can contribute to the wage. These different skill variables can work in concert to provide a more complete profile of the player's ability, and therefore comparing a small set of individual variables might not reflect the full value of the player.

In the post-information era, systems and methods for analyzing football data using computational statistics and pattern recognition have been proposed (Lames et al., 2011). Such

systems have been used for data reduction of players and the ball movement on the field during the game (Feess et al., 2010), automatic detection of player's position (Siegle et al., 2013), interactive football training environment (Jensen et al., 2014), and game tracking (O'Donoghue and Robinson, 2009; Castellano et al., 2014).

In this study we use player skill data to analyze the salary of football players in a quantitative manner as a function of their abilities and skills, and relative to other players. Since the data are multi-dimensional, they are analyzed using machine learning. The goal of the method is to provide a quantitative and objective method for estimating the salary of football players based on their skills and abilities on the field. Such method can be used as an assistive system for contract negotiation, as well as for quantitative estimation of performance per cost and sports economics.

Methods

The methodology used in this study is based on supervised machine learning. That is, the algorithm "learns" from the data samples to deduce a model, and the model is then tested by using other samples that were not used for building the model. These test samples allow to compare the values predicted by the model to the actual values, and measure the accuracy of the model in predicting real samples. In the case of this study, the predicted values are the salaries of the football players, and each sample is represented by a set of variables representing the player's performance and skills as will be described in the next section.

Data

The data used in the experiment are football players' data collected from *sofifa.com*, providing information about the performance and skills of each player, as well as the salary of the players. The salary used in this study is the gross wages paid by the football club as negotiated between the club and the agent of the football player, and excludes other possible income of the player such as commercials, merchandise, etc. Due to the strict rules of football organizations such as FIFA and UEFA, football clubs are required to report the correct salaries to comply with regulations such as the Financial Fair Play (Muller et al., 2012), and¨ due to the careful enforcement of these regulations the reported wages can be considered reliable. Clearly, previous experience has demonstrated some cases in which these regulations were violated in attempt to conceal financial information, but these incidents can be considered exceptions, and the salaries reported by the football clubs can be assumed to be largely reliable.

The data from *sofifa.com* was acquired by applying a web crawler that collected the information of each of the players in the season of 2016. The dataset was originally compiled by football experts, who analyzed the skills of each player for the purpose of simulating the players in video games[1], and consequently for scouting[2]. These data were successfully used for predicting the outcome of football matches (Prasetio, 2016; Shin and Gasparyan, 2014), and have demonstrated to be comparable or better than other sources of football data (Shin and Gasparyan, 2014). The skills are measured as integers in the range of [0,99].

The dataset also includes the salary of each player, which is often used for scouting information and football strategy games such as *Football Manager*. While salary can differ

[1] https://www.theguardian.com/technology/2016/dec/21/fifa-video-game-changed-football

[2]http://www.theguardian.com/football/2014/aug/11/football-managercomputer-game-premier-leagueclubs-buy-players

between leagues, the dataset combines football players from several European football leagues and can therefore vary also based on the league. However, since the Bosman ruling in the mid 1990's (Arnedt, 1998), football players can choose the football club based on the best offers they are given, without the restrictions of nationality, making European football an open labor market (Dejonghe and Van Opstal, 2010).

Table 1 shows the variables used to analyze the overall value of 6,082 players to determine their wages automatically. The variables were collected for 6,082 active football players for which information was available in order to maximize the number of samples in the dataset. The variables include football skills, as well as general physiological variables such as height, weight, and age.

Table 1: Variables used for the salary analysis.

Variable	Description
player_name	Full name of the player
weekly_wage_euros	Weekly salary (€)
Position	Position of the player in the field (e.g., goalkeeper, left winger, defensive midfielder, center forward (Orejan, 2011))
birthday	Birth date
height	Height of the player (cm)
weight	Weight of the player (kg)
preferred_foot	The foot used more commonly to pass or kick the ball (left or right)
crossing	The ability to pass a long ball from the wing, accurately to a teammate inside the penalty box
finishing	The ability to finish the offensive play by scoring a goal from inside the penalty box
heading_accuracy	The ability of the player to reach the ball with their head, and then accurately push it to target
short_passing	Accuracy and speed of short passes to a teammate
volleys	The ability to kick the ball accurately and powerfully while the ball is in the air
dribbling	Kicking the football ball softly ahead with both feet
curve	The technique and ability of a player to kick the ball such that the ball bends while in the air, and accurately reaches a teammate or the opponent's goal
free_kick_accuracy	The likelihood of a free kick made by the player to reach the opponent's goal or a teammate
longpassing	Speed and accuracy of a long kick aiming at passing the ball to a teammate
ballcontrol	How well the player gets control of the ball when making contact with it, and how well they can keep the ball without
acceleration	Ability to speed up within a short amount of time and reach maximum sprint speed
sprint_speed	How fast the player can run when in maximum speed
agility	Ability to make a sudden change of direction and stay in balance
reactions	The time required for the player to respond to certain events such as identifying an opponent running through the defense, identifying a loose ball rejected by the goalkeeper, or responding to a loose ball after a tackle
balance	The ability of the player to stay on their feet in the event of a physical confrontation with another player, or another physical challenges such as jumping
jumping	The elevation and timing of the player when jumping to reach the ball or head hit the ball
stamina	How long/far the player can run in maximum speed before needing to slow down
strength	Player's strength and power, reflected by the ability of the player to overpower an opponent in a physical challenge
shot_power	The player's ability to shot the ball with strength
long_shot	The ability to make an accurate shot from outside of the penalty box
penalties	The ability to score a penalty shot

aggression	The tendency of the player to use strength, leading to more successful tackles yet also leading to more fouls and penalty cards
positioning	The ability of the player to identify open spaces on the field or spots from which gives an advantage for receiving the ball and attacking the opponent's goal
vision	The ability of the player to know where the teammates are positioned on the field, as well as keeping track of the position of the opponent players
marking	The ability to follow an opponent player and prevent them from receiving the ball
standing_tackle	The ability of the player to win the ball in a standing tackle, without committing a foul
sliding_tackle	Ability to win the ball in a sliding tackle
interception	The ability to identify a pass and intercept the ball
gk_diving	Goalkeeper ability to protect the net while in the air
gk_handling	The ability of the goalkeeper to catch and secure the ball
ball_gk_kicking	Goalkeeper ability to make an accurate kick
gk_positioning	Ability of the goalkeeper to find the optimal position while defending the ball
goal_gk_reflexes	The ability of the goalkeeper ability of the goalkeeper to react to a shot and block it

Information from all players analyzed by *sofifa* was used. The dataset included players from 91 football leagues. Table 2 shows the number of teams and number of players from the leagues providing the highest number of players, as well as examples of some other leagues from which less players were used.

Machine learning algorithms

To predict the weekly wage from variables that reflect skills and field performance of a football player, eight different quantitative pattern recognition methods were used. While all of these methods aim at providing the optimal link between patterns of skills and the salary of the players, different supervised machine learning methods perform differently on different machine learning problems and data, and it is often difficult to intuitively select the most effective method without experimenting and comparing the actual performance figures of different methods (Bishop, 2006). Because the salary is a numerical variable, all methods that were selected need to be able to interpolate and predict continuous numerical values, and not be limited to prediction of nominal variables.

The supervised machine learning methods used in the experiments cover several different machine learning paradigms, and include Additive Regression (Friedman, 2002), Decision Table (Kohavi, 1995), Nearest Neighbor with a weighted condition (Aha et al., 1991), K* (Cleary et al., 1995), Locally Weighted Learning with Naive Bayes and Linear regression classifiers (Frank et al., 2002; Atkeson et al., 1997), Random Committee (Seung et al., 1992), and Random Trees (Aldous, 1993).

Additive Regression (Friedman, 2002) is a function-based learning method based on a mathematical equation that smooths the output through using a regression against a linear function, to make the additive regression non-linear function. Decision Table (Kohavi, 1995) is a rule-based learning algorithm that depends on testing a set of data using conditional mathematics, and predict the values of new samples using the results of the training set. Nearest Neighbor with a weighted condition (Aha et al., 1991) is an instance-based method that predicts the outcome of a given situation using the distance equation to find the nearest dataset to predict the results. Another instance-based method is K* (Cleary et al., 1995), which uses the entropy distance from the instances in the training set to predict the results. Locally Weighted Learning is a memory-based learning algorithm, used with naive Bayes or linear regression (Frank et al., 2002; Atkeson et al., 1997), and makes the prediction through the

weighted connection within the data using regression or classification. The Random Committee (Seung et al., 1992) is a meta-learning algorithm that applies different randomly selected number of seeds. Random Tree (Aldous, 1993) is a tree-based method that builds a decision tree using a random selection of attributes in every given dataset. Random Subspace (Ho, 1998) is an ensemble learning method in which a mathematical model creates decision trees, which are combined to make a regression using the posterior probabilities.

Table 2: The number of teams and the number of players from the different leagues in the dataset.

League	**# teams**	**# players**
Serie A (Italy)	20	430
Ligue 1 (France)	20	423
Primeira Liga (Portugal)	18	417
Eredivisie (Netherlands)	18	352
La Liga (Spain)	20	344
Premier league (England)	20	340
Bundesliga (Germany)	18	334
Ekstraklasa (Poland)	16	325
Belgian Pro League (Belgium)	16	257
Championship (England)	24	231
Scottish Premiership (Scotland)	12	222
Swiss Super League (Switzerland)	10	201
Segunda Division (Spain)	22	200
Ligue 2 (France)	20	174
2. Bundesliga (Germany)	18	144
Serie B (Italy)	22	137
I Liga (Poland)	18	93
Super Lig (Turkey)	18	88
League Two (England)	21	86
Scottish Championship (Scotland)	10	84
League One (England)	20	80
Eerste Divisie (Netherlands)	11	70
LigaPro (Portugal)	8	67
Russian Premier League	13	43
Super League (Greece)	10	40
Allsvenskan (Sweden)	11	33
Lega Pro (Italy)	7	31
Danish Superliga (Denmark)	10	25
Austrian Bundesliga (Austria)	9	18
Football National League	4	15
Swiss Challenge League (Switzerland)	2	14
Alka Superliga (Denmark)	1	9
National League (England)	5	7
Cypriot First Division (Cyprus)	4	7
Premier Division (Ireland)	4	4

Each of those algorithms was applied in two different testing strategies. The first used 84% of the samples for training, and the remaining 16% for testing. The second strategy utilized the leave-one-out scheme, using the entire set of 6,082 players. The performance was evaluated by the Pearson correlation between the predicted and the actual weekly wage of the player, as well as the mean absolute error between the actual and predicted wage.

Results

Applying the methods described above to the data described in the Data section using the different machine learning algorithms provided the results shown in Figure 1.

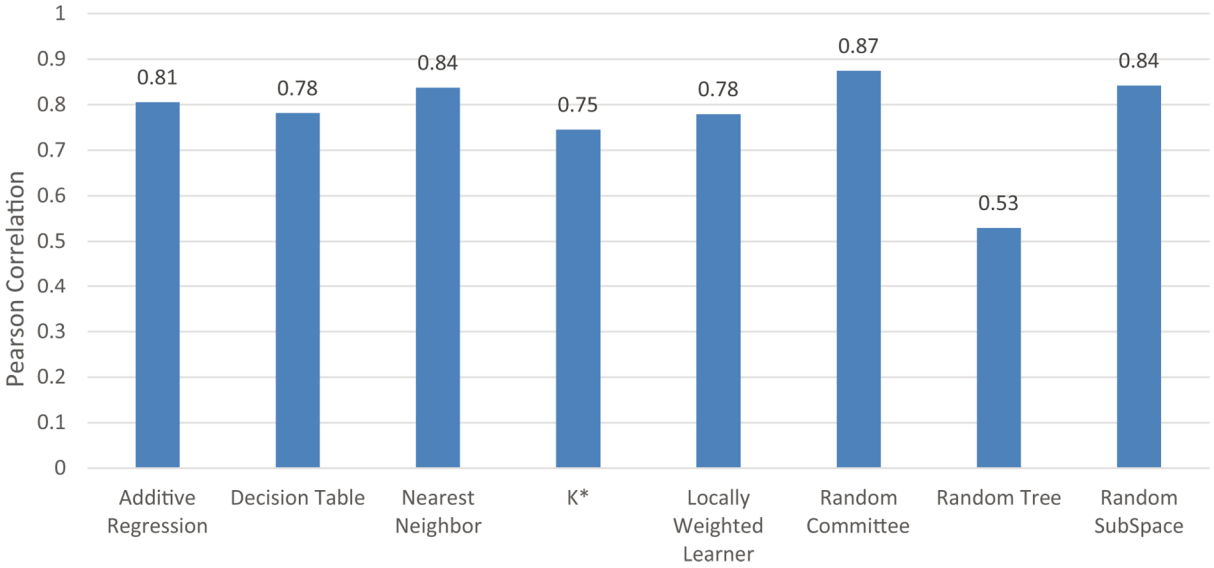


Figure 1: Pearson correlation between the predicted and actual football player weekly wage using ~84% of the samples for training and the remaining ~16% for testing. The probability of all correlations to occur by chance is $p < .001$.

As Figure 1 shows, all machine learning algorithms that were tested show statistically significant correlation between the predicted and actual weekly wage of the players. The weakest algorithm was Random Tree, with a Pearson correlation of 0.53 ($p<.001$). The most effective algorithm for the prediction of the salary is Random Committee, with Pearson correlation of 0.87 between the predicted and actual weekly wage.

Figure 2 shows the mean absolute error between the predicted and actual salaries when using the different machine learning algorithms. As the figure shows, the K-Nearest Neighbor algorithm (K=10) provided the lowest mean absolute difference between the predicted and actual wage, Random Tree provided the highest mean absolute error.

Figure 3 shows the Pearson correlation between the predicted and actual weekly wage using the eight algorithms when using a leave-one-out test strategy.

As the figure shows, all algorithms exhibit a high correlation between the predicted and actual salary, with Random Tree showing a lower correlation than the other algorithms. Figure 4 shows the mean absolute error (in €) between the predicted and actual salary using the leave-one-out testing strategy.

Figure 5 shows the predicted and actual salaries (€) of the 6,082 when using the Additive Regression classifier, and Figure 6 shows the Bland-Altman diagram.

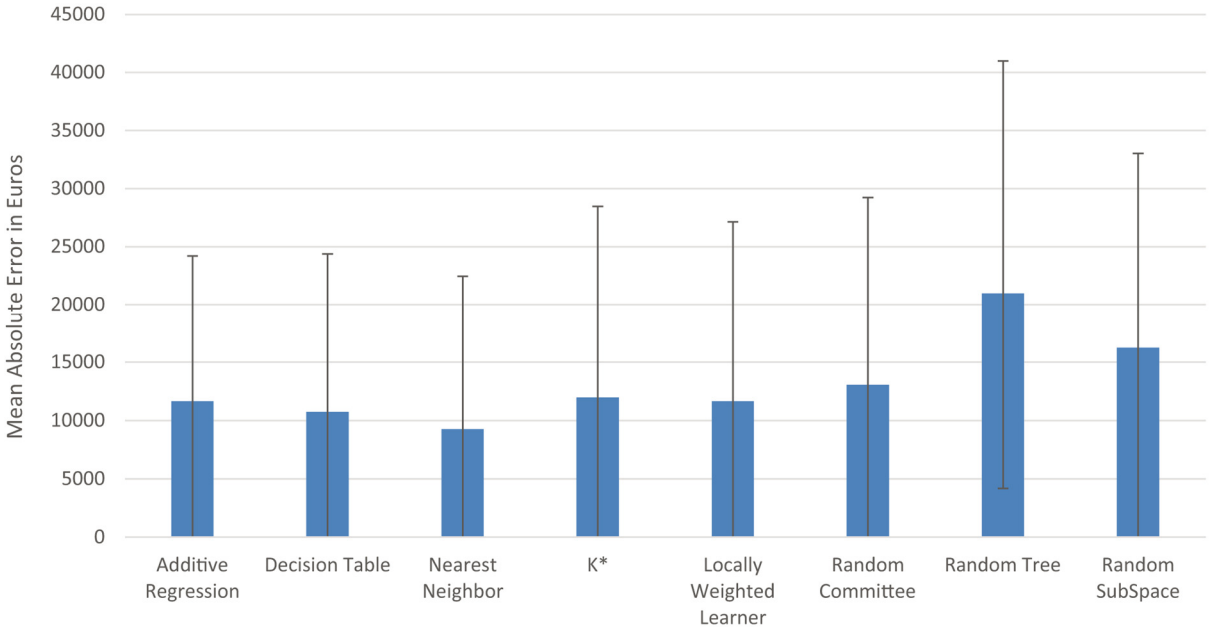


Figure 2: Mean absolute error (€) when using ~84% of the data for training and ~16% for testing. The error bars show the standard deviation of the absolute error.

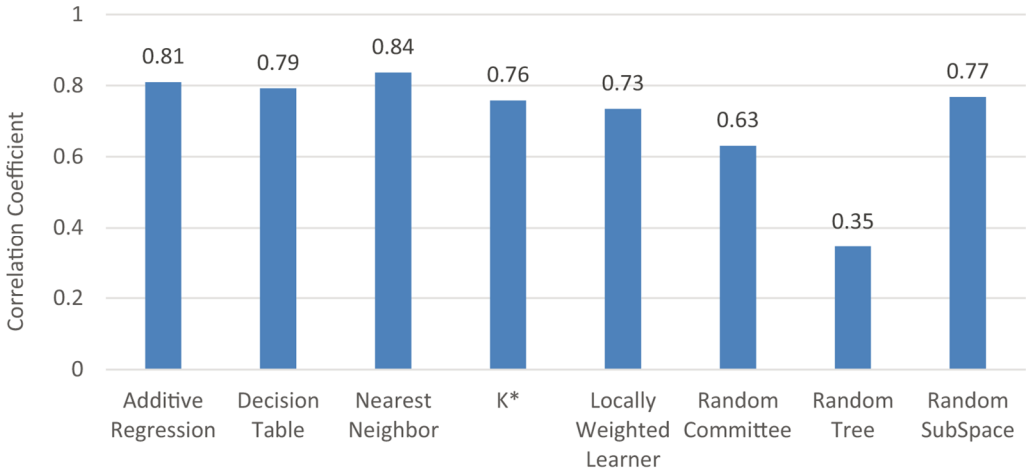


Figure 3: Pearson correlation leave-one-out strategy with the eight machine learning algorithms. The probability of all correlations to occur by chance is $p < .001$.

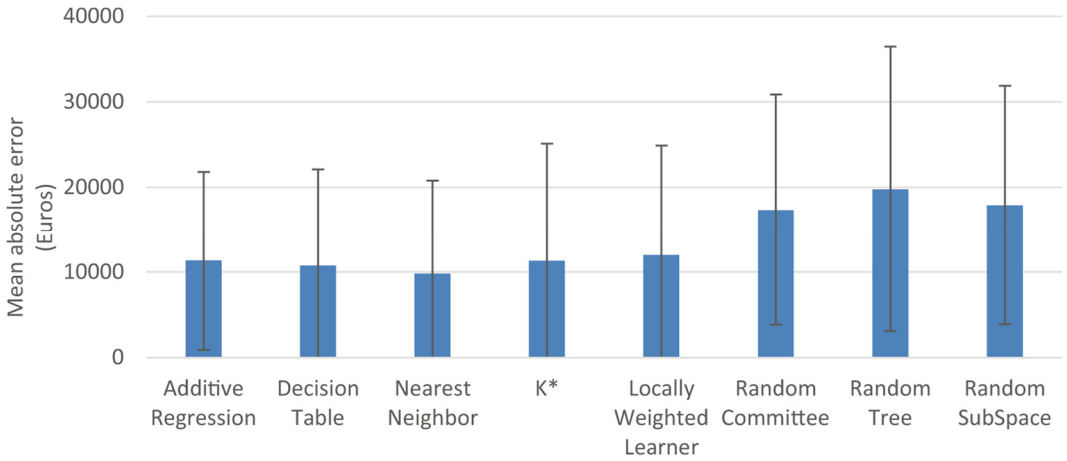


Figure 4: Mean absolute error when using leave-one-out scheme.

Figure 5: Predicted and actual salaries of the 6,082 players when using Additive Regression classifier.

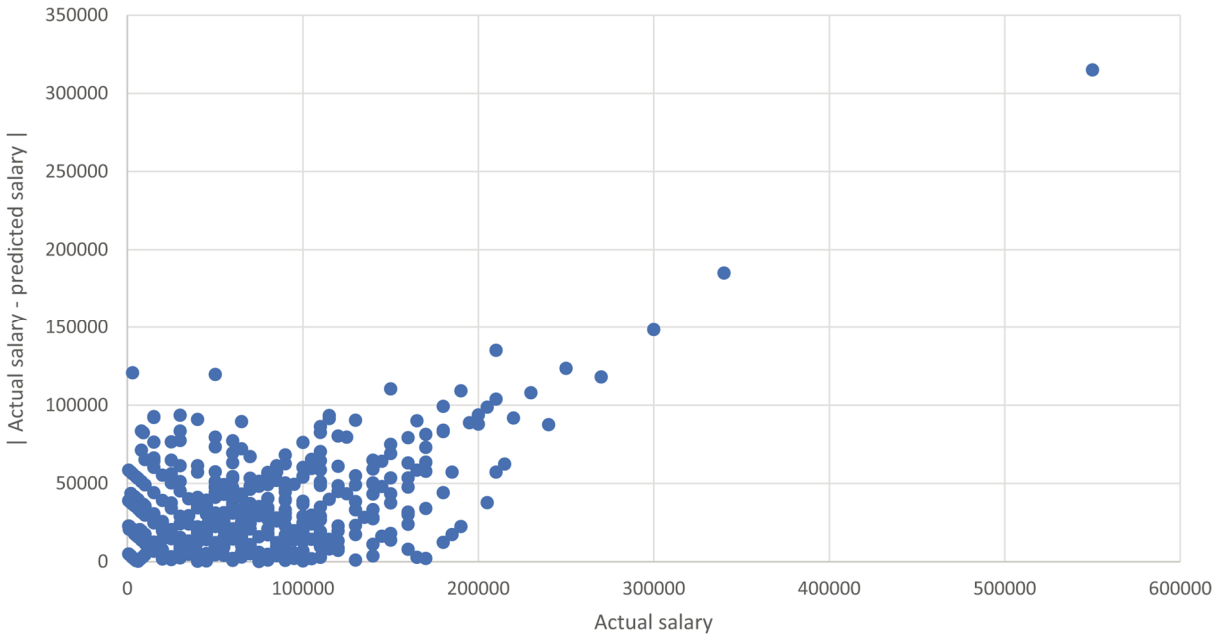


Figure 6: Bland-Altman diagram of the 6,082 players when using the Additive Regression classifier.

Due to the high difference between the salaries of football players shown in Figure 5 and the proportional error shown in the Bland-Altman diagram of Figure 6, we also computed the relative mean error, computed by $E=\frac{\sum_{i=1}^{n}|P_i-T_i|}{\sum_{i=1}^{n}|T_i-\bar{T}|}$, where E is the error, P_i is the predicted wage of player i, T_i is the actual wage of player i, and $\bar{T}$ is the mean wage of all players in the dataset. Figures 7 and 8 show the relative error of the different algorithms when separating the data to training and test sets and when using the leave-one-out strategy, respectively.

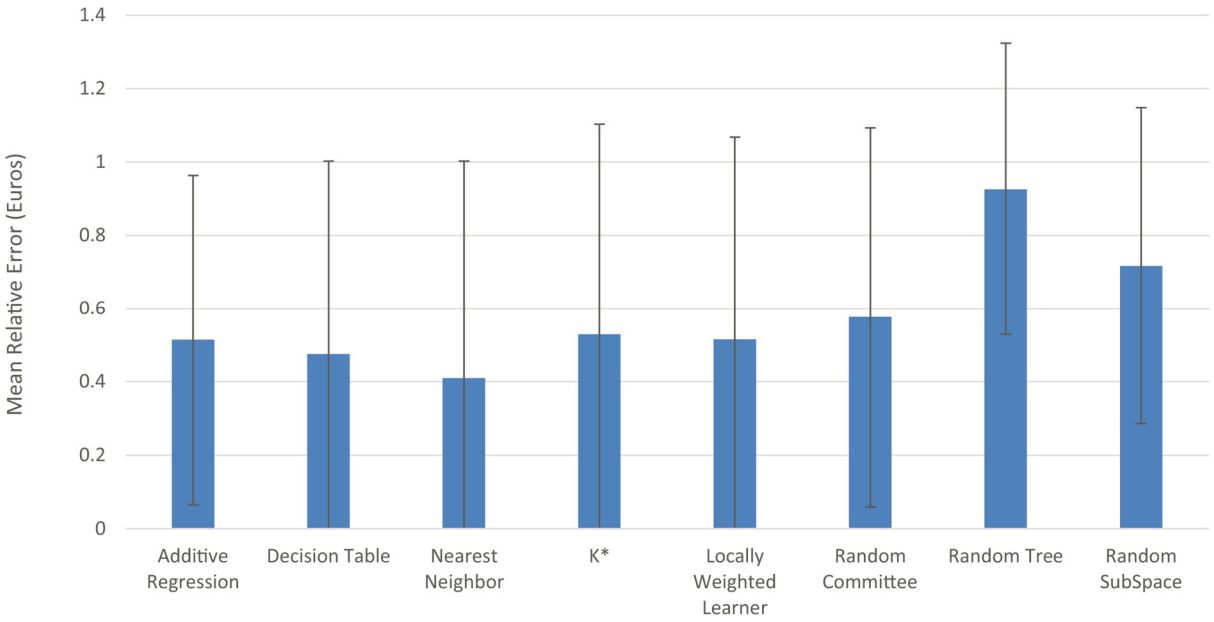


Figure 7: Relative error when using ~84% of the data for training and ~16% for testing.

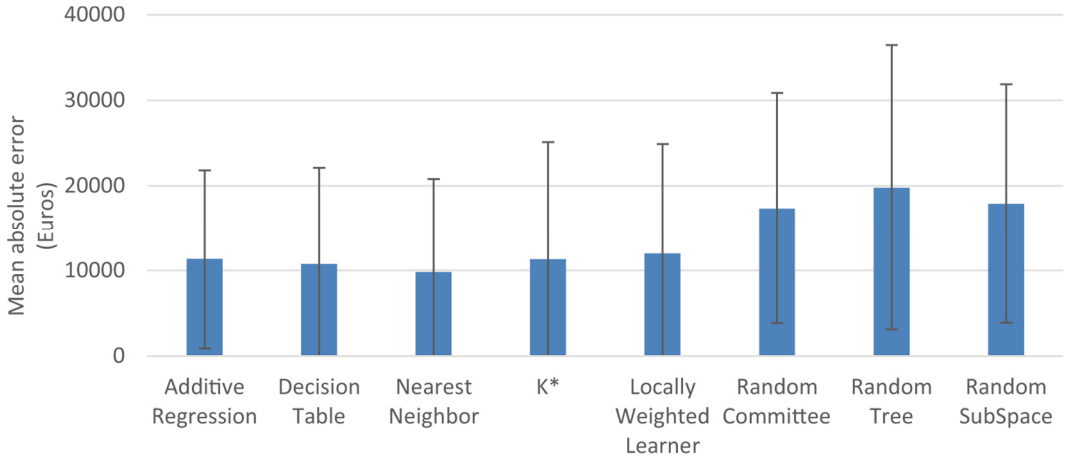


Figure 8: Relative error when using the leave-one-out test strategy.

To identify the skills that have the highest impact on the salary, the Relief Attribute Evaluation (Robnik-Sikonja and Kononenko, 1997) was used. The variable selection method worksˇ by repetitively selecting samples and evaluating the weight of the feature by comparing the nearest samples of the same class and of the other classes. The variables that were selected are listed in Table 3.

Predicting the wages of individual football players

The method was also applied to predict the salaries of individual football players. Table 4 shows the predicted salaries of several known football players with the highest salaries when using Additive Regression model such that the training set was the 6,081 other players, and the single test sample was the football player of which the salary is predicted.

Table 4 shows that the algorithm was able to automatically identify the highest paid football players by merely analyzing their performance figures and skills. For instance, of the 100 top earning football players in 2016, 61 were among the 100 top predicted salaries. The probability of 61 randomly selected players to be among a set of 100 players selected from a set of 6,082 players is ($p < 10^{-5}$). However, in many cases it also shows differences between the actual and

predicted salary of the players. For instance, for players such as Karim Benzema and Lionel Messi the weekly wages have been predicted approximately 50% less than their actual weekly wage. Other players with the highest difference between predicted and actual salary are Petr Cech, with actual weekly salary of €160K and predicted salary of ~€123K, and Oscar, who earn €140K per week while his predicted wage is ~€67K. That difference can be explained by the monopsony rent as discussed in the Introduction, and the contention that football superstars provide their club not merely with their performance on the field, but also with increasing fan base, merchandise sales, income from broadcasting rights, etc (Garciadel Barrio and Pujol, 2007; Frick, 2006). On the other hand, N'Golo Kante has a predicted salary of ~€106K while his actual weekly wage is €50K, the predicted wage of Timo Horn is ~€61K while his actual wage is €25K, and Tim Howard has a predicted wage of ~€87K while his actual wage of €40K is over 50% lower. The additive regression classifier was used for its ability to better handle edge cases, while its overall performance (Pearson correlation of 0.81) is very close to the performance of the Nearest Neighbor algorithm (0.84) as shown in Figure 1.

Table 3: The top skills ranked automatically by their relevance using the Relief attribute evaluation method.

Skill rank	Skill
1	reactions
2	finishing
3	standing tackle
4	position
5	ball control
6	sliding tackle
7	interceptions
8	heading accuracy
9	long passing
10	vision

Analysis by different football leagues

Compensation for football players can vary between leagues, as salary standards for one league can depend on the popularity of the game in that country, the strength of the economy, the football play style in the league, etc, that might lead to different salaries for different players, or preference for different skills. For instance, the German Bundesliga might prefer strength and toughness while the Spanish La Liga might reward players with higher technical skills. To normalize for the different leagues, the same experiment was repeated for each of the eight leagues with the highest number of players: Bundesliga (Germany), La Liga (Spain), Serie A (Italy), Premier League (England), Ligue 1 (France), Ekstraklasa (Poland), Primeira Liga (Portugal), and Eredivisie (Netherlands).

Table 5 shows the Pearson correlation between the computed and actual salaries when using the different machine learning algorithms and the leave-one-out testing strategy. The table shows that unlike the experiment with the entire dataset of players, the Nearest Neighbor algorithm does not provide the highest correlation between the predicted and actual salaries when the data points (players) are separated into leagues. That can be explained by the fact that Nearest Neighbor is an instance-based algorithm that its performance depends heavily on the size of the training set (Dasarathy, 1994), and therefore its performance is affected substantially from the smaller size of the dataset. The table also shows that in all leagues the correlation between the predicted and actual salaries is largely similar given a certain classifier,

with the exception of the Polish top division, showing a far weaker correlation between the salaries and the skills of the players.

Table 4: Predicted salaries when using Additive Regression classifier and the actual weekly wages of football players with the highest salaries in the dataset.

Players Name	**Recent Salary**	**Predicted Salary**
Lionel Messi	550,000	235,026
Cristiano Ronaldo	340,000	154,766
Luis Suarez	300,000	151,567
Neymar	270,000	151,940
David De Gea	250,000	126,314
Mesut Oezil	240,000	152,314
Angel Di Maria	230,000	121,935
Sergio Busquets	220,000	128,050
Luka Modric	215,000	152,625
Thomas Mueller	210,000	152,687
Marco Reus	210,000	106,111
Gareth Bale	210,000	152,687
Robert Lewandowski	210,000	74,861
Sergio Aguero	210,000	106,111
Robin van Persie	210,000	152,687
Andres Iniesta	210,000	106,111
Edinson Cavani	205,000	167,242
Karim Benzema	205,000	106,151
Jerome Boateng	200,000	106,191
Diego Costa	200,000	111,948
James Rodriguez	195,000	106,231
Gerard Pique	190,000	167,482
Gonzalo Higuain	190,000	80,681
Manuel Neuer	185,000	167,562
Eden Hazard	185,000	127,776
Jordi Alba	180,000	135,845
Henrik Mkhitaryan	180,000	96,001
Thibaut Courtois	180,000	80,714
Paul Pogba	180,000	167,642
Ivan Rakitic	180,000	96,878
Javier Pastore	180,000	167,642
Nicolas Otamendi	180,000	167,642
Isco	170,000	88,532
Mario Goetze	170,000	167,803
Philippe Coutinho	170,000	167,803
Vincent Kompany	170,000	112,223
David Luiz	170,000	112,223
Yaya Toure	170,000	96,912
Cesc Fabregas	170,000	106,431
Antoine Griezmann	170,000	135,968
Diego Godin	170,000	106,371
Pedro Rodriguez	165,000	74,971

Table 5: Pearson correlation between the actual and predicted salaries using the different algorithms in the different leagues. The probability of all correlation to occur by chance is $p < .001$.

Algorithm	Serie A	Primeira Liga	Premier League	Ligue 1	Eredivisie	Bundesliga	La Liga	Ekstraklasa
Additive Regression	0.76	0.70	0.78	0.79	0.61	0.78	0.74	0.34
Nearest Neighbor	0.53	0.50	0.53	0.55	0.39	0.56	0.63	0.26
K*	0.68	0.59	0.72	0.68	0.60	0.74	0.75	0.35
Locally weighted learning.	0.60	0.59	0.66	0.53	0.63	0.69	0.46	0.09
Random Committee	0.77	0.73	0.75	0.78	0.66	0.79	0.77	0.49
Random Sub Space	0.70	0.57	0.71	0.70	0.58	0.66	0.72	0.41
Decision Table	0.66	0.56	0.74	0.73	0.66	0.85	0.74	0.44
Random Tree	0.51	0.39	0.43	0.55	0.38	0.57	0.64	0.29

The differences between the leagues can also be reflected by the variables that are the most relevant for the determination of the salary. The variable selection was done by using the Relief Attribute Evaluation (Robnik-Sikonja and Kononenko, 1997).

Table 6 shows the highest ranked attributes (football skills) in different football leagues. While the table shows good agreement between the leagues, some variables are more relevant to the determination of the salaries in some leagues and are ranked lower in other leagues. For instance, quick reactions (*reactions*) is a major attribute in all leagues, except for Portugal's Primeira Liga, where reactions are also relevant to the salary, but the variable is ranked lower compared to the other leagues. In the Spanish top football division the ability to finish a play close to the goal (*finishing*) or to kick the ball before it hits the ground (*volleys*) is ranked high compared to the other leagues. Heading accuracy is more relevant to the salary in Spain, Germany, and Portugal, compared to the other top leagues. In the Polish Ekstraklasa the variable that has the strongest impact on the salary is the team of the player. That league also prioritize other variables that do not have strong link with the wage in the other leagues such as *jumping*, *weight*, and *acceleration*. The preferred foot have a higher impact on the wage in England and France, while interceptions are ranked low in the Netherlands.

Results by field positions

Different positions might require different sets of skills, and therefore the players can be separated by their field position to allow a more consistent analysis of the skills required for each position. Table 7 shows the frequencies of the different field positions in the dataset. Tables 8, 9 and 10 show the correlation, mean absolute error, and relative absolute error, respectively, for the different field positions.

As the tables show, the Decision Table algorithm provided the best performance, while the lowest performance was observed with Random Tree. All field positions showed a good correlation, and no position is characterized by a more predictable salary compared to the other positions. The P values of the Pearson correlation coefficients computed using all methods were lower than .001.

Table 6: The top variables ranked automatically by their relevance using the relief attribute evaluation method in the different football leagues.

Variable rank	Serie A	Primeira Liga	Premier League	Ligue 1	Eredivisie	Bundesliga	La Liga	Ekstraklasa
1	reactions	position	reactions	standing tackle	reactions	reactions	reactions	team
2	position	standing tackle	preferred foot	reactions	position	penalties	volleys	penalties
3	sliding tackle	finishing	interceptions	preferred foot	short passing	vision	finishing	jumping
4	ball control	heading accuracy	jumping	marking	crossing	sliding tackle	heading accuracy	reactions
5	interceptions	free kick accuracy	long passing	interceptions	ball control	heading accuracy	preferred foot	marking
6	marking	team	sliding tackle	free kick accuracy	standing tackle	interceptions	standing tackle	standing tackle
7	standing tackle	penalties	standing tackle	jumping	sliding tackle	crossing	ball control	free kick accuracy
8	vision	reactions	marking	penalties	marking	standing tackle	interceptions	positioning
9	finishing	vision	gk kicking	balance	long passing	short passing	vision	ball control
10	gk diving	crossing	short passing	sliding tackle	volleys	ball control	gk positioning	weight
11	crossing	Long passing	vision	volleys	gk handling	gk handling	marking	gk positioning
12	positioning	Sliding tackle	gk handling	heading accuracy	standing tackle	gk kicking	sliding tackle	acceleration
13	aggression	marking	ball control	ball control	acceleration	gk reflexes	gk kicking	hot power
14	dribbling	curve	gk positioning	crossing	heading accuracy	gk diving	gk reflexes	gk reflexes
15	long passing	interceptions	heading accuracy	Position	jumping	volleys	gk diving	dribbling
16	gk positioning	long shots	gk reflexes	finishing	finishing	marking	ball control	volleys
17	short passing	ball control	gk diving	short passing	dribbling	finishing	positioning	gk diving
18	gk reflexes	gk diving	dribbling	weight	positioning	sprint speed	balance	sliding tackle
19	Heading accuracy	dribbling	height	gk kicking	gk kicking	free kick accuracy	position	height
20	volleys	gk reflexes	balance	long passing	penalties	positioning	weight	vision

Table 7: Frequency and number of players in each field position in the dataset.

Field position	# players	Frequency
Forwards	1,678	0.28
Midfielders	2,258	0.37
Defenders	1,634	0.27
Goalkeepers	512	0.08

Table 8: Pearson correlation between the predicted and actual salaries based on position on the field. The probability of all correlations to occur by chance is $p < .001$.

	Forwards	**Midfielders**	**Defenders**	**Goalkeepers**
Additive Regression	0.80	0.83	0.83	0.85
Nearest Neighbor	0.65	0.69	0.69	0.70
K*	0.77	0.79	0.80	0.69
Locally Weighted Learning	0.66	0.66	0.73	0.72
Random Committee	0.80	0.82	0.83	0.83
Random SubSpace	0.69	0.69	0.74	0.64
Decision Table	0.85	0.82	0.87	0.81
Random Tree	0.58	0.46	0.62	0.54

Table 9: Mean absolute error and standard deviation between the predicted and actual salaries based on position on the field.

	Forwards	**Midfielders**	**Defenders**	**Goalkeepers**
Additive Regression	11,834±10,414	11,503±10,519	9,492±9,781	11,043±9,511
Nearest Neighbor	16,095±14,956	14,917±13,661	12,851±11,104	14,447±11,385
K*	11,685±13,322	11,798±12,871	9,320±10,095	13,297±13,711
Locally Weighted Learning	16,378±16,949	16,053±16,746	12,554±12,917	14,935±14,064
Random Committee	14,823±13,668	14,433±13,514	11,805±11,072	14,585±13,114
Random SubSpace	20,117±19,255	19,551±18,897	16,205±16,562	20,551±18,883
Decision Table	10,405±10,084	10,562±10,235	8,101±7,867	10,217±9,221
Random Tree	17,589±17,858	18,263±18,691	13,873±15,594	17,296±17,149

Table 10: Mean relative error between the predicted and actual salaries based on the position on the field.

	Forwards	**Midfielders**	**Defenders**	**Goalkeepers**
Additive Regression	0.51	0.51	0.48	0.48
Nearest Neighbor	0.69	0.66	0.67	0.63
K*	0.50	0.52	0.49	0.58
Locally Weighted Learning	0.71	0.71	0.66	0.65
Random Committee	0.64	0.64	0.62	0.64
Random SubSpace	0.87	0.87	0.85	0.90
Decision Table	0.45	0.47	0.42	0.45
Random Tree	0.76	0.81	0.73	0.76

Overpaid and underpaid players

As the previous experiments show, some players are underpaid by their clubs, while other players are compensated in a fashion that is not proportional to their performance on the field. To identify skills that differentiate between underpaid and overpaid players, the dataset was sorted by the absolute difference between the predicted and actual salary, as well as by the predicted and actual relative salary.

When observing the absolute difference, the most underpaid players are Harry Kane

(Tottenham Hotspur, England), Bernardo Silva (Monaco, France), and Granit Xhaka (Arsenal, England), while the most overpaid among the 6,802 players based solely on their football skills are Lionel Messi (F.C. Barcelona, Spain), Angel Di Maria (Paris Saint-Germain, France), and Robin van Persie (Fenerbahe, Turkey). As mentioned earlier, the salary is predicted based on the performance and skills of the players, and does not take into account marketing and fan base considerations.

The relative difference show that the most underpaid players are Julian Weigl (Borussia Dortmund, Germany), Jefferson Lerma (Levante UD, Spain), and Bartlomiej Dragowski (Serie A, Italy), and the most overpaid players are Gregory Bourillon (Angers, France), Darko Lazovic (Genoa, Italy), and Sergey Krivets (Wisa Pock, Poland). Tables 11 and 12 show the most underpaid and most overpaid players, respectively, computed by using a decision table.

Table 11: The actual and predicted salary of the most overpaid players in 2016.

Player	**Actual wage (€)**	**Predicted wage (€)**
Lionel Messi	550,000	243,333
Angel Di Maria	230,000	58,974
Robin Van Persie	210,000	61,428
Ivan Rakitic	180,000	61,556
Nicolas Otamendi	180,000	61,556
Edin Dzeko	165,000	53,873
Philippe Coutinho	170,000	60,552
Isco	170,000	60,552
David Luiz	170,000	61,599
Sergio Busquets	220,000	115,342
Edinson Cavani	205,000	103,298
Lucas Moura	130,000	32,385
Cristian Zapata	110,000	12,814
Alessio Cerci	115,000	18,062
Gareth Bale	210,000	115,559
Karim Benzema	205,000	115,668
Mario Balotelli	100,000	11,125
Sokratis Papastathopoulos	150,000	61,684
Juan Cuadrado	150,000	61,684
Laurent Koscielny	150,000	61,684

To identify the skills that differentiate between underpaid and overpaid football players, the 100 most underpaid players and the 100 most overpaid players were separated, and the Linear Discriminant Analysis (Bishop, 2006) was performed to identify the skills that have the highest LDA scores, indicating that they provide the strongest separation between the overpaid and underpaid players. The vast majority of the skills had LDA score close to 0, indicating that single skills do not have substantial impact on the salary, and the salary is determined by a combination of skills. However, some skills showed statistically significant difference. Table 13 show the mean and LDA score of the skills with the highest LDA scores.

Table 12: The actual and predicted salary of the most underpaid players in 2016.

Player	Actual wage (€)	Predicted wage (€)
Bernardo Silva	10,000	119,907
Harry Kane	15,000	119,798
Granit Xhaka	70,000	171,176
Timo Horn	25,000	119,581
Francisco Alcacer	80,000	170,588
Gelson Dany Batalha Martins	30,000	119,472
Renato Sanches	35,000	119,364
Paul Nardi	25,000	104,548
Andrea Pirlo	50,000	129,500
Lavyin Kurzawa	50,000	119,038
Joao Mario	50,000	119,038
Daniel Amartey	6,000	69,758
Iuri Medeiros	2,000	64,973
Diogo Jota	3,000	64,947
Leroy Sane	3,000	64,947
Riechedly Bazoer	3,000	64,947
Jefferson Lerma	1,000	62,318
Niklas Suele	4,000	64,921
Corentin Tolisso	45,000	104,409

Table 13: Means and standard error of the mean of the skills that have the highest LDA scores between the 100 most overpaid and 100 underpaid players. The table also shows the t-test P value of the difference between the means.

Skill	Mean underpaid	Mean overpaid	LDA score	t-test p
Reactions	78.4±0.4	71.5±0.7	0.67	.00
Vision	68.0±1.2	61.5±1.5	0.10	.00
Acceleration	73.0±1.4	67.7±1.8	0.08	.02
Agility	71.8±1.2	66.9±1.5	0.07	.01
Sprint speed	72.9±1.2	68.4±1.4	0.06	.01
Balance	68.9±1.3	64.3±1.5	0.05	.02

When ranking the players by the most underpaid and overpaid players measured by the relative error between the predicted and actual salaries, the skills that had the highest LDA score are listed in Table 14. Although the LDA scores were lower compared to using the absolute error, the consistency shows that some skills can differentiate between overpaid and underpaid players. For instance, overpaid players tend to be stronger than underpaid players, while the underpaid players, on average, outperform the overpaid players in their reactions, agility, balance, acceleration, and sprint speed.

Table 14: Means and standard error of the mean of the skills that have the highest LDA scores between the 100 most overpaid players and 100 underpaid players when measured by the relative error.

Skill	Mean underpaid	Mean overpaid	LDA score	t-test p
Reactions	68.6±0.6	65.2±0.8	0.12	.00
Strength	64.9±1.3	70.5±1.5	0.10	.00
Agility	68.5±1.3	64.4±1.5	0.04	.04
Balance	66±1.2	62.9±1.4	0.03	.10
Acceleration	68.5±1.3	65.4±1.5	0.02	.12
Sprint speed	68.8±1.3	66.3±1.4	0.02	.19

Conclusion

Estimating the salary of a football player is a task determined by negotiation between the football clubs and the agents of the football player. In addition to the economical aspects of team budget, the salary of a football player also affects his or her performance, and also the performance of the other players of the team (Torgler and Schmidt, 2007). Here we described a quantitative method that is based on the skills and performance figures of the player, in relation to the football economy reflected by the wages earned by other players. The method proposes an objective scale that is based on the football economy market, and therefore can optimize the football club resource allocation.

Clearly, the salary of a football player is affected by numerous factors that are not directly related to performance or skills. For instance, a player liked by the fans can have higher compensation to reflect merchandise and ticket sales. For instance, a player such as Lionel Messi receives a weekly wage of €550,000 given that part of it is sponsorship, merchandise sales, image rights, and other incomes outside the scope of wins or performance on the field. Using pattern recognition, Messi's predicted salary computed based on the performance and wages of the current football players in the major football leagues would have been about 50% less, although according to the method he would still be the world's highest paid player.

The observation that the most effective results were obtained using the instance-based nearest neighbor algorithm show that salaries of players similar in their set of skills also tend to be more similar in their salaries. That is, players with similar skill sets tend to be more similar in their salaries compared to players who have different strengths. For instance, two players who are equally good in both skills *A* and *B* will have a more similar salary compared to two players such that one is stronger in *A* but weaker in skill *B*, and the other is stronger in *B* while weaker in *A*. That is in agreement with the relatively low performance figures of the tree-based algorithms, in which the decisions based on the skills would have expected to outperform the other algorithms if one strength balancing a weakness would be a more dominant factor when determining the player's salary. When the players were separated into leagues, however, the performance of the nearest neighbor algorithm was substantially weakened, as nearest neighbor is an algorithm that its efficacy depends heavily on the size of the training set (Dasarathy, 1994).

Analysis of the most underpaid and the most overpaid players shows that overpaid players are physically stronger than underpaid players and the difference is statistically significant, but the overpaid players are inferior in their reactions, vision, acceleration, agility, and balance compared to underpaid football players.

The data used in this study are taken from football expert analysis performed for the initial purpose of using the data for football video game simulators. The ability of football simulation video games to reflect the football reality has been advancing rapidly in the past two decades (Markovits and Green, 2016). For instance, star football players have donated their body motions to allow reliable reflection of soccer player motions during the game, wellknown narrators and football analysts donated their voice, and virtual replications of dozens of football stadiums have been created. Naturally, substantial efforts have been invested in analyzing the skills and performance of the football players to allow reliable simulation of the games that reflects the reality of football matches. These datasets demonstrated efficacy in predicting the outcomes of football match, and were equal or better than other football data (Prasetio, 2016; Shin and Gasparyan, 2014).

The machine learning-based observation that the salaries of football superstars are not proportional to their performance on the field is aligned with the economical analysis showing that the salary of football superstars are biased by the monopsony rent (Garcia-del Barrio and Pujol, 2007). The recruitment of football superstars is also important for branding, and can be used as a marketing strategy of football clubs, and therefore the efforts of a football club to recruit them is not necessarily driven solely by football strategy aiming at achieving the optimal achievements on the field (Kase et al., 2007). That "superstar effect" has been shown to affect crowd attendance (Bryson et al., 2014), and also contributes to the financial success of the football club in addition to its sporting success (Rohde and Breuer, 2016).

The method proposed in this paper is driven by quantitative analysis, and the salary is determined solely by the measured skills and performance of the athlete. Using a quantitative method as a baseline for football player salaries might also have a behavioral impact on the team, as it has been shown that salary inequality has a negative impact on football player performance, and the negative effect gets larger as the difference between the salary of the player and the salary of the other players in the team increases (Torgler et al., 2006; Torgler and Schmidt, 2007). Knowing that the salaries are determined such that an objective quantitative method is used as baseline might have an impact on the performance of the athlete, who knows that the key for determining the salary is equal to all players. Such methods can be used as a baseline to simplify the negotiation process, and determine a uniform salary scale that given its objective quantitative nature might be perceived as more consistent compared to the current agent-club negotiation.

Acknowledgements

We would like to thank the anonymous reviewers for the insightful and constructive comments that helped to improve the manuscript. This research was funded in part by NSF grant IIS-1546079.

References

Aha, D. W., Kibler, D., & Albert, M. K. (1991). Instance-based learning algorithms. *Machine Learning*, *6*(1):37–66.

Aldous, D. (1993). The continuum random tree III. *The Annals of Probability*, 248– 289.

Arnedt, R. B. (1998). European union law and football nationality restrictions: the economics and politics of the bosman decision. *Emory International Law Review*, *12*, 1091.

Atkeson, C. G., Moore, A. W., & Schaal, S. (1997). Locally weighted learning for control. In *Lazy learning* (pp. 75-113). Springer Netherlands.

Bishop, C. M. (2006). Pattern recognition and machine learning. *Machine Learning*, *128*,1–58.

Bryson, A., Rossi, G., & Simmons, R. (2014). The migrant wage premium in professional football: a superstar effect? *Kyklos*, *67*(1), 12–28.

Castellano, J., Alvarez-Pastor, D., & Bradley, P. S. (2014). Evaluation of research using computerised tracking systems (amisco R and prozone R) to analyse physical performance in elite soccer: A systematic review. *Sports Medicine*, *44*(5), 701–712.

Cleary, J. G., Trigg, L. E., et al. (1995). K*: An instance-based learner using an entropic distance measure. In *Proceedings of the 12th International Conference on Machine learning*, *5*, 108–114.

Dasarathy, B. V. (1994). Minimal consistent set (mcs) identification for optimal nearest neighbor decision systems design. *IEEE Transactions on Systems, Man, and Cybernetics*, 24(3), 511–517.

Dejonghe, T. & Van Opstal, W. (2010). Competitive balance between national leagues in european football after the bosman case. *Rivista di Diritto ed Economia dello Sport*, *6*(2), 41–61.

Feess, E., Gerfin, M., & Muehlheusser, G. (2010). The incentive effects of long-term contracts on performance-evidence from a natural experiment in european soccer. Technical Report, Mimeo: Berlin.

Frank, E., Hall, M., & Pfahringer, B. (2002). Locally weighted naive bayes. In *Proceedings of the Nineteenth Conference on Uncertainty in Artificial Intelligence*, 249–256.

Frick, B. (2006). Salary determination and the pay-performance relationship in professional soccer: Evidence from germany. *Sports Economics After Fifty Years: Essays in Honour of Simon Rottenberg. Oviedo: Ediciones de la Universidad de Oviedo*, 125–146.

Frick, B. (2007). The football player's labor market: Empirical evidence from the major european leagues. *Scottish Journal of Political Economy*, *54*(3), 422–446.

Frick, B. (2011). Performance, salaries, and contract length: empirical evidence from german soccer. *International Journal of Sport Finance, 6*(2), 87.

Friedman, J. H. (2002). Stochastic gradient boosting. *Computational Statistics & Data Analysis*, *38*(4), 367–378.

Garcia-del Barrio, P., & Pujol, F. (2007). Pay and performance in the spanish soccer league: who gets the expected monopsony rents. Technical report, University of Navarra, Spain

Garcia-del Barrio, P., & Pujol, F. (2009). The rationality of under-employing the bestperforming soccer players. *Labour*, *23*(3), 397–419.

Giulianotti, R. (2012). *Football.* Wiley Online Library.

Ho, T. K. (1998). The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *20*(8), 832–844.

Jensen, M. M., Grønbæk, K., Thomassen, N., Andersen, J., and Nielsen, J. (2014). Interactive football-training based on rebounders with hit position sensing and audio-visual feedback. *Intentional Journal of Computer Science in Sport*, *13*(1), 57–68.

Kase, K., De Hoyos, I. U., Sanchis, C. M., & Breton, M. O. (2007). The proto-image of real madrid: implications for marketing and management. *International Journal of Sports Marketing and Sponsorship*, *8*(3), 7–28.

Kohavi, R. (1995). The power of decision tables. In *European Conference on Machine Learning*, 174–189.

Lames, M., McGarry, T., Nebel, B., & Roemer, K. (2011). Computer science in sportspecial emphasis: Football (dagstuhl seminar 11271). *Dagstuhl Reports*, *1*(7).

Markovits, A. S., & Green, A. I. (2017). FIFA, the video game: a major vehicle for soccer's popularization in the United States. *Sport in Society, 20*(5-6), 716-734.

Muller, J. C., Lammert, J., & Hovemann, G. (2012). The financial fair play regulations of˙ uefa: An adequate concept to ensure the long-term viability and sustainability of european club football? *International Journal of Sport Finance*, *7*(2), 117.

O'Donoghue, P. & Robinson, G. (2009). Validity of the prozone3 r player tracking system: A preliminary report. *International Journal of Computer Science in Sport*, *8*(1), 37–53.

Orejan, J. (2011). *Football/Soccer: History and tactics*. McFarland. Jefferson, NC, USA.

Prasetio, D. (2016). Predicting football match results with logistic regression. In *International Conference on Advanced Informatics: Concepts, Theory And Application*, 1–5.

Robnik-Sikonja, M. & Kononenko, I. (1997). An adaptation of relief for attribute estimation in regression. In *Machine Learning: Proceedings of the Fourteenth International Conference*, 296–304.

Rohde, M. & Breuer, C. (2016). Europes elite football: Financial growth, sporting success, transfer investment, and private majority investors. *International Journal of Financial Studies*, *4*(2), 12.

Seung, H. S., Opper, M., & Sompolinsky, H. (1992). Query by committee. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, 287–294.

Shin, J. & Gasparyan, R. (2014). A novel way to soccer match prediction.Technical Report, Stanford U., CA. USA.

Siegle, M., Stevens, T., & Lames, M. (2013). Design of an accuracy study for position detection in football. *Journal of Sports Sciences*, *31*(2), 166–172.

Torgler, B. & Schmidt, S. L. (2007). What shapes player performance in soccer? empirical findings from a panel analysis. *Applied Economics*, *39*(18), 2355–2369.

Torgler, B., Schmidt, S. L., & Frey, B. S. (2006). Relative income position and performance: an empirical panel analysis.

Wicker, P., Prinz, J., Weimar, D., Deutscher, C., & Upmann, T. (2013). No pain, no gain? effort and productivity in professional soccer. *International Journal of Sport Finance*, *8*(2), 124.