Application of Regression Tree in Salary Estimation of Team Players in Final of European Football Championship 2016

Yulisman Zega¹, Bagus Sartono², Agus Mohamad Soleh³
¹²³Department of Statistics, Bogor Agricultural University, INDONESIA

ABSTRACT
Regression tree is one of nonparametric data exploration methods that can be used to observe relationship between continuous response variables and explanatory variables that are large and complex. The explanatory variables considerably affect the response are variables that appear as a separator or partition. This study aimed to analyze the factors affecting the estimation of team player salary in final of European Football Championship 2016. The results illustrated that after trimmed by Minimum Complexity Cost rule, R²CV values obtained was 0.798 with classification accuracy level of 72.09%. Explanatory variables that partitioned were passing/header/shooting, became the dominant variable in explaining the salary estimation of team player in final of European Football Championship 2016.

Keywords— European football championship 2016 team, regression tree, salary estimation

I.  INTRODUCTION

Football is a sport that uses feet. In the game, the sport is played by two opposing groups called teams. Football is a team sport that reflects the culture of the player. Style of the football game flows continuously and requires concentration in solving the problem spontaneously and naturally [1]. The modern football firstly conceived and spread worldwide by the British people and continues to grow now a day.

According to a FIFA survey in 2006, approximately 243 million people play professional football, which accounts for 4.1% of the world population. It indicates the development of football as a global sport continues to grow along with the times [6]. The high salary of players per week is assessed on skill of the player itself. Skill of players intended is the number of performance during senior career, number of goals during senior career, loyalty to the club, passing/heading, speed/reflex, durability/air capability, Shooting/foot skills.

The method used to investigate the most influencing factors on the salary per week of football players is the method of regression tree. Regression tree method is a nonparametric method used to analyze the relationship between the response variables and the explanatory variables including nominal, ordinal and continuous data. Regression tree can be a good choice for analysts which gain fairly accurate and fast results, but it needs more time and special when uses traditional methods. In addition, the regression tree is very helpful when uses a lot of variables and it is also able to identify the variables that interacted each other [5].

II.  METHODOLOGY

Data
The data used in this study were secondary data from football data (Country Teams of European Football Championship) in 2016 [8]. The data collected from 430 players from 24 European countries were analyzed. Response variable (Y) observed was salary per week (pounds), while the independent variables considerably affect the response variable were:

Analysis Method
The analysis consisted of several stages as follows:
1. Descriptive analysis of response variable and the explanatory variables to determine the general conditions.
2. Perform analysis of regression tree
   a. Determine the maximum tree
   b. The tree was iteratively pruned into a smaller tree sequence and nested
   c. Choose the best tree of the sequence using the sample test estimate or cross validation estimate

The best way to do the validation and pruning is to use the data group of explanatory variables with cross validation method. 10- fold cross validation)
will produce a final tree [10]. Estimates of the quadratic mean error of cross validation is formulated with:

$$R_{CV}(T_k) = \frac{1}{N} \sum_{i=1}^{N} [Y_i - \hat{Y}(k)]^2$$

with $R_{CV}(T_k)$ is the square mean error optimal tree trimmed , $N$ is the number of observations in a node $k$ , $y_i$ is the value of the response variable , and $\bar{Y}(k)$ is the average of the node $k$ . best tree is $T_k0$ are:

$$R_{CV}(T_{id})= \min R_{CV}(T_i)$$

with $R_{CV}(T_{id})$is the square mean error best optimal regression tree [3] . The core of minimum complexity pruning cost is cutting ties weakest regression tree.

3. Determine the response estimation for each end node
4. Value of response estimation on each group of observations made was the response average
5. Interpret the results obtained from the regression tree with analysis process using R software.

### III. RESULT

European Football Championship 2016 hosted by France was attended by 24 countries including Spain as the defending champion. The data were secondary data of 430 football players from all teams in European Football Championship 2016 with salaries weekly calculated in pounds unit (£).

**TABLE 1**

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean</th>
<th>St dev</th>
<th>Min.</th>
<th>Max.</th>
</tr>
</thead>
<tbody>
<tr>
<td>salary</td>
<td>£50808.2</td>
<td>£48412.2</td>
<td>£220</td>
<td>£288000</td>
</tr>
</tbody>
</table>

Table 1 showed that the average of player salary was £50808.2 with standard deviation of £48412.2. The range between the maximum and the minimum value indicates that there are many outliers in the data.

**Description of player salary by Country**

Every country has best players that were selected from some clubs either elite club or new developing club, and based on identity of players, so that player recruitment depends on the consideration of coach from each country by considering the skill of player during joining the match at the club they played for. This selection was not only in terms of skill but also in terms of discipline within the club, as well as other considerations aspects based on the rules of player recruitment in country itself. Coaches prefered players who have already been popular and on the rise, besides they were easier to play, they also have undoubted talents and skills. The description of player salaries based on their origin country as follows:

**TABLE 2**

<table>
<thead>
<tr>
<th>Country</th>
<th>Mean</th>
<th>Country</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Germany</td>
<td>£ 88.358</td>
<td>Sweden</td>
<td>£ 41.226</td>
</tr>
<tr>
<td>Spain</td>
<td>£ 87.814</td>
<td>Switzerland</td>
<td>£ 38.493</td>
</tr>
<tr>
<td>France</td>
<td>£ 79.324</td>
<td>Russia</td>
<td>£ 34.900</td>
</tr>
<tr>
<td>England</td>
<td>£ 72.100</td>
<td>Austria</td>
<td>£ 32.740</td>
</tr>
<tr>
<td>Belgium</td>
<td>£ 61.850</td>
<td>Slovakia</td>
<td>£ 30.114</td>
</tr>
<tr>
<td>Portugal</td>
<td>£ 50.804</td>
<td>Rep. of Ireland</td>
<td>£ 28.878</td>
</tr>
<tr>
<td>Cezch rep</td>
<td>£ 46.750</td>
<td>Rumania</td>
<td>£ 27.420</td>
</tr>
<tr>
<td>Italy</td>
<td>£ 46.533</td>
<td>Iceland</td>
<td>£ 22.400</td>
</tr>
<tr>
<td>Swiss</td>
<td>£ 45.242</td>
<td>Northern Ireland</td>
<td>£ 19.581</td>
</tr>
<tr>
<td>Poland</td>
<td>£ 43.829</td>
<td>Ukraine</td>
<td>£ 16.850</td>
</tr>
<tr>
<td>Croasia</td>
<td>£ 43.300</td>
<td>Albania</td>
<td>£ 13.120</td>
</tr>
<tr>
<td>Turkey</td>
<td>£ 41.424</td>
<td>Hungary</td>
<td>£ 10.753</td>
</tr>
</tbody>
</table>

Table 2 showed that the country with the highest average salary was Germany of £88,358 with standard deviation of £55,878, followed by Spain of £87,814 with standard deviation of £57,528 and French of £79,324 with standard deviation of £35,324, while that the country with the lowest average salary was Hungary of £10,753 with standard deviation of £10,753. The high salary of players from Germany, Spain and France due to the presence of many players who came from the big clubs such as Barcelona, Real Madrid, Chelsea, Bayern munchen, Juventus and arsenal that always play in Union of European Football Associations (UEFA) Champions League and other championships or leagues, so these clubs become famous and as a consequence, every player was expensively paid to be more competent during the tournaments held. In constrast, players of the Ukrainian, Albania and Hungary came from local clubs which are not well-known yet and just follow the match between clubs in local region.

Based on the geographical position, the European continent is a very large continent which divided into five parts: Western, Central, Eastern, Northern and Southern Europe. European Football Championship is followed by all countries in Europe which started firstly by following the qualification phase. European Football Championship is held once in four years. In recruiting players, every country has focused to their ability to pay each player. Western European countries have an average salary greater than eastern europe. Germany is part of Central Europe but more inclined to western Europe, while Spain, France and the UK are part of Western Europe. On the other hand, Ukraine, Albania and Hungary are more likely part of Eastern Europe. So, Western Europe has higher average salary than eastern Europe.
Figure 1: Country’s Geochart based on salaries

Figure 1 showed that the highest salary are marked with red color then decreases until a dark green color. Each country adapts to colors according to the salary average.

In football team, each player has their position and duties corresponding to their characteristics and skill. The position is crucial in achieving a victory, placing the right person in corresponding position to their expertise. That position in the match also determines the amount of salary in every match. Placing a player in a certain position can be determined by assessing their ability during the match. The match record of each player has taken by coach as consideration for selecting the name that will join as the Country’s team. In European Football Championship 2016, every country puts players in accordance with its domain expertise. These positions were goalkeeper (48 players or 11.16%), back position (120 players or 27.91%), the center position (143 players or 33.26%) and the front position (119 players or 27.67%).

<table>
<thead>
<tr>
<th>Position</th>
<th>Mean</th>
<th>St dev</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Goalkeeper</td>
<td>£42.165</td>
<td>£48.838</td>
<td>£900</td>
<td>£200.000</td>
</tr>
<tr>
<td>Back</td>
<td>£48.354</td>
<td>£48.354</td>
<td>£2.600</td>
<td>£197.000</td>
</tr>
<tr>
<td>Middle</td>
<td>£52.234</td>
<td>£49.324</td>
<td>£1.200</td>
<td>£256.000</td>
</tr>
<tr>
<td>Defender</td>
<td>£55.056</td>
<td>£55.576</td>
<td>£220</td>
<td>£288.000</td>
</tr>
</tbody>
</table>

Table 3 showed that position at the front has the highest salary of £ 55,056, then the center position of £ 52,234, and then the back position of £ 48.354 and last position as a goalkeeper of £ 42.657. Striker or commonly referred as the forward player has a huge responsibility to win every match. Stricker is the most prominent player where his position greatly influence the success of goals. Striker needs agility, speed, accuracy of shooting, heading ability and good ball control, so that the front position as the striker has a greater salary than other positions. The position of the center also has an important role as the forward player that connects between forward and back player. Their duty is to bring the ball to the front during attacking and prevent the opponent to attack the defense area. Back position plays role to protect the goalkeeper when take the ball, prevents the opponent to attack and blocks the shot from opponent. Back position is for the player who is strong, hardworking and fast to anticipate the opponent movement. The duty of goalkeeper is to save the defense area or the goal from the opponent's attack, usually handled by player who is tall, big and has a speed in catching the ball. In addition, a goalkeeper must also has a strong kick as his duty to adress the ball to the team. The ability possessed by each player varies according to their talents and tenacity in performing exercise. The existing talent is honed and developed to get a positive value. Passing skill is one of the very basic techniques in playing football. Pass correctly is key to a team in controlling the game due to the entire match of football is the giving and receiving feedback. Heading skill is a technique in putting the ball in a perfect target. This skill greatly affects a person to achieve success in the match. Armed with the skill of passing, heading, shooting and the placing exact position in a match will gain a very satisfactory result, where the placement of the player based on their skill is very influential.

Regression Tree of Player Salary Variables

Data analysis in determining the regression tree was conducted in two stages. The first stage was to determine the maximum tree of the overall explanatory variables exist, then pruning conducted in order to obtain the optimal tree. Regression tree method produced a model that is simple and easy to interpret. The model resulted was based on influential variables as the identifier of forming a node. First explanatory variable partitioned was passing / heading / shooting that appeared on the first node. This variable was considered as the variable that most influence to the amount of football players salary, so the variable was the dominant variable.

The early stages resulted the tree that produced a maximum size of 13 nodes. Complexity Cost Minimum Rules, which is one method of trimming to get the optimal tree using a 10-fold cross-validation with the smallest RCV value (x-val Relative Error) of 0.798, obtained the optimum tree (Figure 2)
The tree had nine terminal nodes with observation classifications accuracy level of 91.56% with R-square of 0.576 (Figure 3) and the explanatory variable that partitioned was passing/heading/kick, so the variable was the dominant variable in explaining the salary estimation of players in European Football Championship 2016 team. Other variables that arose were the speed/reflex, age, the performance number for senior career, loyalty to the club, number of goals during senior career.

IV. CONCLUSION

This study shows that the estimation of team player salary in European Football Championship 2016 is affected by the passing / heading / kick and speed/reflex. In more detail can be described as follows:

1. Based on regression tree trimming, estimation of player salaries were:
   a. The highest salary was £144,900 with the ability to passing/heading/shooting <7.75, speed/reflex ≥ 6.75 age of <30.5 years old loyalty to the club ≥ 6.75
   b. The Lowest salary was £21,440 with passing/heading/kick <7.75, speed/reflex <6.75, age ≥ 31.5 years

2. Tree regression showed that the skill and age greatly affect the high estimation of player salary.

3. The good regression tree is a tree that does not have a lot of outliers with high classification accuracy and high R-square

REFERENCES