FIFA - Player Analysis and Club Standing Predictions

Anand Krishnamoorthy, Anugraha Venugopal, Devanshi Deswal, Manaswini Nagaraj

There is an avid interest in statistical methodology in sports. Within the sports business, football is a lucrative industry associated with more than half of the global population. Understanding player skills is an integral part of the analysis of these methods. In this project, we aim to analyze football players in consensus with FIFA(Fédération Internationale de Football Association) records in search of answers that can help increase the performance of individuals or clubs in their respective leagues and even enable prediction of the game.

The dataset "FIFA 19 Complete Player Dataset" is publicly available on Kaggle and dwells over 18,207 players spanning over 89 attributes about personal details, evaluation of their finesse at 26 field positions, and quantification of 34 skill matrices. Approximately seven percent of total players are goalkeepers, who play an entirely different game in comparison to their teammates. Hence, there is missing data for goalkeepers in positional dimension- CF(Center Forward), LB(Left Back), RM(Right Midfielder), etc. since they cannot substitute outfield players. This contributes to approximately 50% of the missing data. The remaining missing data is a result of omitted observations for 48 football players. On further research it was found that these are players who have played less than 100 matches in FIFA, hence do not have accurate statistics value- Positioning, Strength, Balance, etc. which otherwise would have led to poor predictions results on the model.

Using exploratory analysis, we propose to evaluate a player as well as the club standing via skill and positional variables. Using p/f statistics we intend to test the hypothesis that the valuation of a player is based on age and position. Further, we test whether age reduces the rating of a player, i.e. hurt the player's overall rating, and answer questions like "Should the team renew the contract of players or should they try to recruit new players to increase the chances of winning" and in general predict the players net worth. The key idea is that the managers can use data on the player's statistics to predict the best team formation for a high win probability. The correlation among predictors would be analyzed (using plots, Pearson/ spearman coefficients) followed by a dimensionality reduction – using lasso/ PCA. The missing values would be taken care of by imputation techniques. We then build linear models, regularized linear models and regression trees in combination with cross-validation for estimation.

We further intend to model the probability where potential and talent meets. There are young players with more talent and a low rating. They might be willing to move to a bigger club to unleash their potential. Modeling player statistics for different positions and skillsets would help identify a similar player and ideal replacement for substitutions during matches. Handling highly reputed players is important for team management. Therefore, a weighted average prediction method is recommended where more emphasis is on predicting correctly the high reputed players. Our analysis should identify the best attributes using PCA or using forward/backward selection techniques. To separate and classify the players we would use SVM(Support Vector Machine) and make modeling predictions based on decision trees, bagging, boosting and neural nets. We could use cross-validation as model evaluation and the model would require feature scaling and label encoding of the attributes. For business evaluation, we interpret the results from the models and identify what leads to a particular prediction using LIME/ SHAP.

Finally, grouping the club players to identify the top 20 clubs and the distribution of their players in these clubs to set a basis for the establishment of match tactics. Evaluating the defense, center, attack, and goalkeeping of the team can help identify team strength and weakness. The manager may use the evaluations to build a defense from the back or use the defensive system as a form of attack against their opponent. Here we build several models like trees, random forests and tune the hyper-parameters. With answers from all these questions, the team managers are better equipped to march their teams to victory.