BUILDING THE A-TEAM

ISYE 6501 INTRODUCTION TO ANALYTICS MODELING

COURSE PROJECT

PREPARED BY: SERENA TAY

CASE STUDY: SCISPORTS

- Purpose: Leveraging data analytics for player recruitment in association football
- Objective: Finding the player(s) that creates an optimal combination of players in a current team

"The best teams aren't those with the best players, but the best combination of players"

Case study link: https://www.sas.com/en_us/customers/scisports.html

DATA SELECTION

Splitting the data: Given that a football team is comprised of players playing different positions, it is reasonable to split the players up into general groups namely

- Goalkeepers
- Defenders: right backs, left backs, center backs
- Midfielders: Right winger, left winger, center midfielder
- Attackers: Strikers, Forwards

With players segregated in these groups, it is easier to analyze potential purchases for a specific position.

Assumption: Objective is not to build the team from scratch but to bolster a team with player(s) of a certain position to ensure the optimal combination of players is achieved

Note: Subset of data above is for Step 2 onwards

Sourcing the data: Can be sourced from various third-party data collectors such as OptaJoe and WhoScored

STEP 1: SELECTING THE COMBO

Given {data} Players and their statistics for each team

Use {model} K-Means Clustering

Using this unsupervised approach, the aim is to identify clusters of football clubs based on the type players that make up the team and their individual statistics. This is to identify teams with a similar set of players, style of play and formation.

Advantages: This method allows a quick analysis on the available data and removing potential bias in selecting predictors that would exist in a supervised approach

Disadvantages: There is limited control from the modeler's perspective. For example, one would not be able to specific a column of data as the target data.

Note: Correlation clustering can be a potential alternative to the generic clustering model to see if there is a potential optimum number of clusters the software might suggest.

To {Result} identify teams with similar sets of players

Using the clusters formed, one would detect the cluster where the club (club who is interested in purchasing a player) is located in and use clubs in that same cluster for the next analysis.

Note: Clustering is used in this case to detect clubs in the same cluster as it would allow us to identify clubs across different leagues that comprised a similar combination of players to perform further analysis on the next step

STEP 2: SELECTING THE PERFORMANCE PREDICTORS

Given {data} Players within cluster and within specific playing position to be recruited
For example, if a certain club is looking for a defender, the subset of defenders from the cluster is used.

Use {model} LASSO regression
Assuming that data for all variables/predictors are readily available from the dataset, LASSO may be the ideal choice for this next step, which is to identify the predictors that correlate with players with higher overall team performance.

Advantages of LASSO:
The reason is mainly there will not be any cost associated for selecting a predictor over the other depending on what the model choose as the final sets of variables. The case would be different if one is required to acquire or data and additional information for predictors selected if they are not readily available.

Disadvantages of LASSO:
On the other hand, as Dr. Sokol mentioned in the lecture, it is not entirely clear how LASSO weighs the predictive power of one variable over the other, so for analysis purposes, it might not provide us with a justification of why some variables are excluded and why the rest are retained.

To {result}
Although variable selection is not a must, it would definitely help simplify the model by reducing the number of variables and potentially provide insights on which variables/factors that influences/attributes that are important in identifying higher performers.

Please see **Other Considerations** for further discussion on the response variable for this analytical model.

STEP 3: SELECTING THE BETTER TEAM PLAYER

Given {data}

1. Players within the same cluster and plays in the position that the club is looking for recruits (see Step 1)
2. Predictors selected in Step 2

Use {model} Bayesian Modelling

Using the highest predictor(s) selected in Step 2, perform a Bayesian modelling on all combination of players. Using results of all the models created for each pair of players, rank the players based of their overall results.

Advantages: Using the Bayesian modelling over all combination of players allow a more comprehensive one-to-one comparison with other players

Disadvantages: Depending on the number of players in the cluster of the same position, there may be a significant number of combinations that needs to be tested which requires cost of time and analysts.

To {result}

The result of the modelling above would create an overall ranking of players of a certain position based on their comparative performance against all the players within the same cluster. The club and analysts would then assess their performance and their estimated market value to find option(s) that are good for value.

Note: Please note that the modelling can be done on one single strongest predictor or a few to get a more balanced results. However, this would increase the number of models by n-fold for n number of additional predictors. If this alternative is pursued, a linear regression would need to be created based on the predictors chosen to identify an appropriate weighting of each predictors for the overall performance result.

OTHER CONSIDERATIONS

The current modelling techniques noted in this deck are used to only find players from a comparable combination of players in a team, there are potential more in-depth areas for exploration such as the follows:

- Step 1 alternative: Instead of clustering teams with a similar combination of players, one may consider clustering individual players with similar statistics compared to a player in the current team to initiate the analytics process. The idea is to find potential recruits with similar playing styles/performance in this clustering analysis.

- Step 2 Response Predictor selection: The exact response predictor has been kept ambiguous intentionally as this variable is dependent on the respective position one is recruiting for and the preferences of scouts and football managers on which metric they want to use to “measure” a player’s performance.

OTHER CONSIDERATIONS

The assumption made for this project is to find a player that is similar to a player in the team to ensure suitable replacement and compatibility. Hence there is another potential area for analysis that revolves around a team aiming to improve their team's quality in order to get to the next level in competition by finding a good combination of players. Possible analysis includes the following:

- Alternative analysis: Explore clubs with similar team formation and combination of players that are performing at a higher level (i.e. higher win to loss record, points achieved or etc.) to seek out any insights on what key factors influenced their superior performance