Examining Substitution Decision Making in the Canadian Premier League Using Machine Learning

Anonymous

10/02/2020

Abstract

One of the most influential changes that a manager can make in a game of soccer is the substitution. Using a game-by-game dataset from the Canadian Premier League (provided by StatsPerform), this work attempts to understand what statistics (visual statistics like turnovers and underlying statistics like expected goals) coaches may be making substitutions based off. We employ random forest modeling to see what variables are important to the decision-making process. This information and analysis can be replicated with similar datasets to assist owners and general managers in the coaching recruitment process.

1 Introduction

Substitutions have played a vital role in determining soccer matches. Of the past three major international tournament finals, two were 1-0 victories where a substitute scored the winning goal. However, the priority with substitutions isn't always to bring an attacker on to try and win a game. It can be to change formational shape or to sure up defenses. The primary focus of this work, however, is looking at why players are brought off. Understanding this can help to unlock what coaches are trying to accomplish with their substitutions.

1.1 Dataset

The dataset featured in this work is from the 2019 Canadian Premier League season. The 2020 season is not used in this dataset because the teams were not limited to three substitutions. Each row within the dataset represents a player's statistics within a certain match.

Goalkeepers will be removed from the initial dataset, as the decision to take off a goalkeeper is, for the most part, entirely down to injury reasons (Within the dataset, only 1 goalkeeper(s) was substituted off). The dataset will also remove players who are subbed on, as we can't build the decision model on players who can't be decided on. The dataset does not specify whether players were taken off because of injury or not. To account for this, we will remove any players who were taken off before halftime. Overall, there are 1972 datapoints that will be used. They must also be seperated into their respective teams.

All statistics will be used as rate statistics (the statistic divided by the minutes played in a match). This is to make sure that player statistics are not lowered because they were subbed off.

1.2 Related Work

The inspiration for this paper was Bret Myers' "A Proposed Decision Rule for the Timing of Soccer Substitutions". While the work done for that paper was focused on the timing of substitutions and on

creating a strategy for managers, this paper focuses more on the statistics behind them and on just observing managerial behavior.

2 Analysis

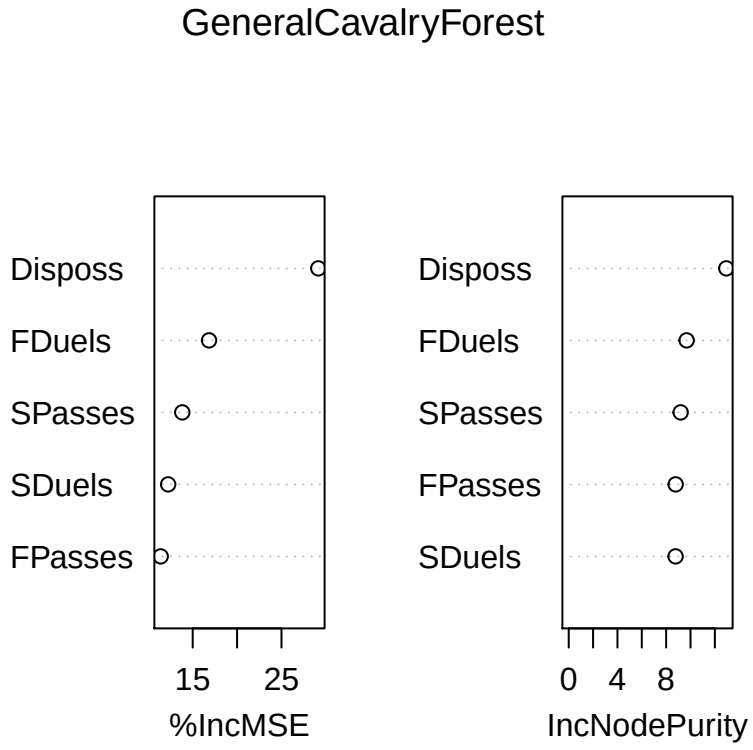
For the analysis, we are going to train models for every team based on four categories: General, Attackers, Midfielders, Defenders. We will use more general statistics and have all players within a team be used to train the model. For the latter three categories, we will only use players within the specified positions and focus on statistics that are relevant to the position.

2.1 General

For the general dataset, we will build the random forest model based on five variables (which have been abbreviated for convenience):

- Successful Duels Per Minute -> SDuels
- Failed Duels Per Minute -> FDuels
- Successful Passes Per Minute -> SPasses
- Failed Passess Per Minute -> FPasses
- Total Times Dispossesed Per Minute -> Disposs

There are two methods of assessing how much a variable is important to the success of a random forest model. The first is the mean decrease accuracy, which assesses how the accuracy of a model would decrease as a result of removing a certain variable. The second is mean decrease gini (also known as increased node purity), which assesses how much a certain varible plays into observations being decided correctly. The figure below illustrates these two methods using the random forest model built on data from Cavalry FC.



The total times that a player is dispossesed per minute seems to be the most influencing factor on the head coach of Cavlary FC, Tommy Wheeldon Jr., when making a substitution. However, we have only focused on five statistics that are not positional specific, so our analysis is limited.

2.2 Attackers

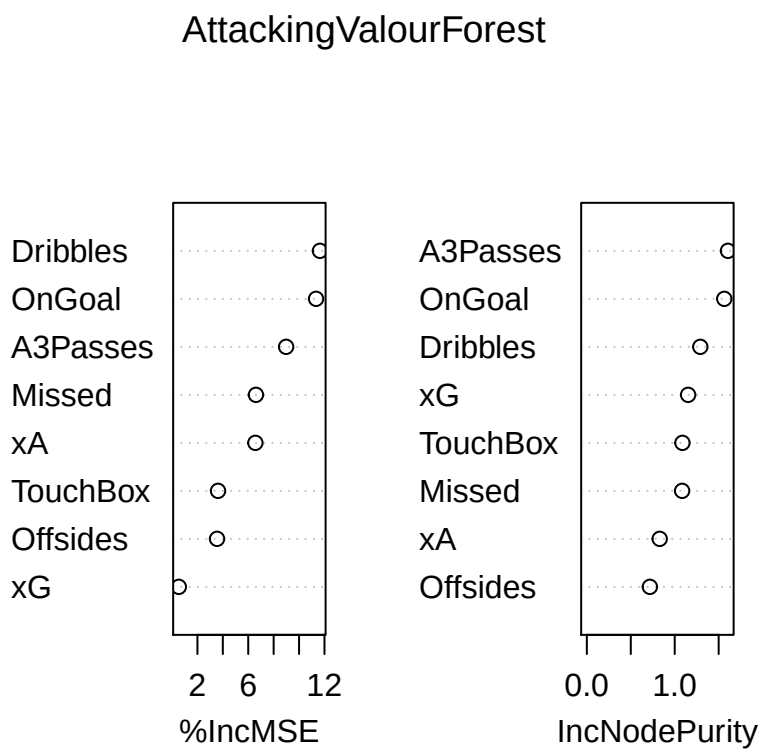
Focusing on attackers now, we will have access to four positions that were specified within the dataset:

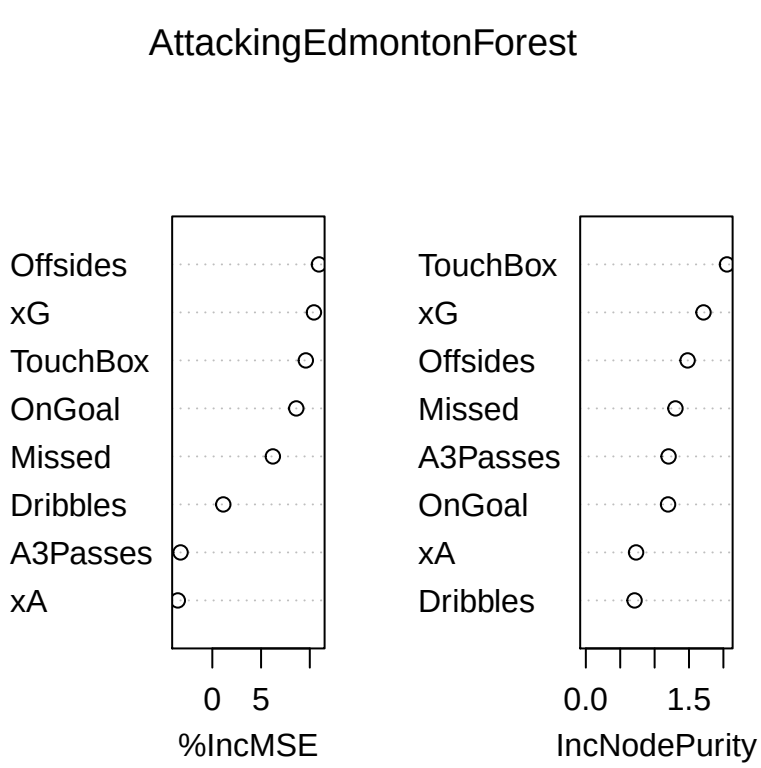
- Centre Forward
- Left Winger
- Right Winger
- Second Striker

The statistics (and abbreviations) that will be used for this updated dataset are:

- Expected Goals Per Minute -> xG
- Expected Assists Per Minute -> xA
- Shots On Goal Per Minute -> OnGoal
- Missed Shots Per Minute -> Missed
- Touches in the Opponent's Box Per Minute -> TouchBox
- Successful Passes in the Attacking Third Per Minute -> A3Passes

For this analysis, we will focus on two teams, FC Edmonton and Valour FC, and see what may be influencing the head coaches' decisions.





Rob Gale of Valour FC seems to be basing his attacking decisions based off of shot accuracy (high emphasis on shots on target) and keeping hold of the ball in dangerous positions (high emphasis on accurate passes in the final third and successful dribbles). Jeff Paulus of Edmonton FC looks to be basing his decisions based on off the ball-movement (high emphasis on offsides and touches in the opponent's box) and getting quality chances (high emphasis on xG).

2.3 Midfielders

We now shift to midfielders, where we will have access to seven positions within the dataset:

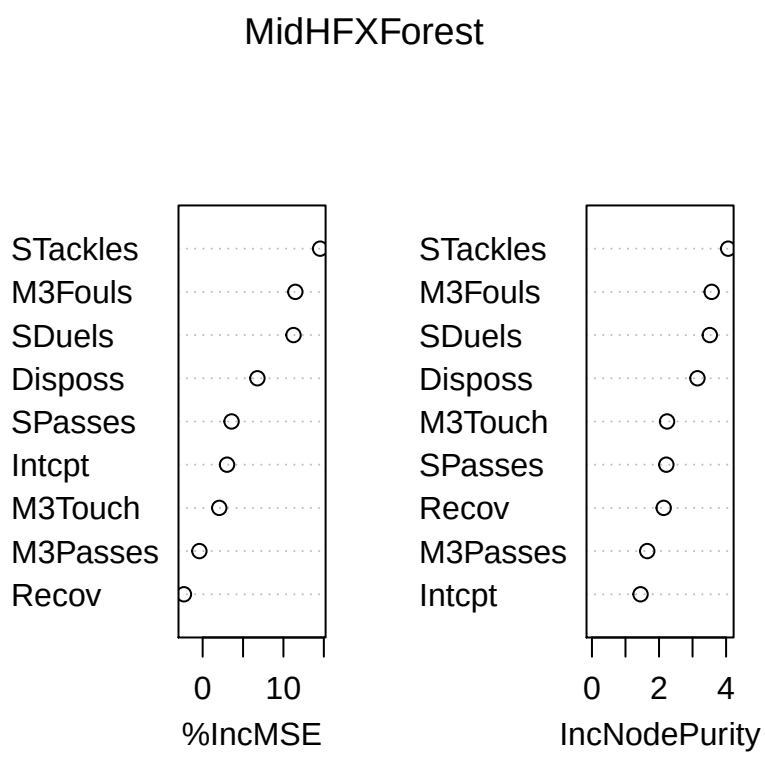
- Central Midfielder
- Centre Attacking Midfielder
- Defensive Midfielder
- Left Attacking Midfielder
- Right Attacking Midfielder
- Left Midfielder
- Right Midfielder

The statistics (and abbreviations) that will be used for this updated dataset are:

- Successful Duels Per Minute -> SDuels
- Successful Passes Per Minute -> SPasses
- Successful Tackles Per Minute -> STackles
- Interceptions Per Minute -> Intcpt
- Recoveries Per Minute -> Recov

- Passes Completed in the Middle Third Per Minute -> M3Passes
- Touches in the Middle Third Per Minute -> M3Touch
- Fouls in the Middle Third Per Minute -> M3Fouls

In this section, we will focus on HFX Wanderers and their head coach Stephen Hart.



A lot of the metrics that appear on the top of both of the variable importance charts are centered around duels and tackles, rather than passes or interceptions. Steven Hart may be looking for players who aren't winning their duels and are committing too many fouls in the middle of the field when he is deciding who to take off in the midfield.

2.4 Defenders

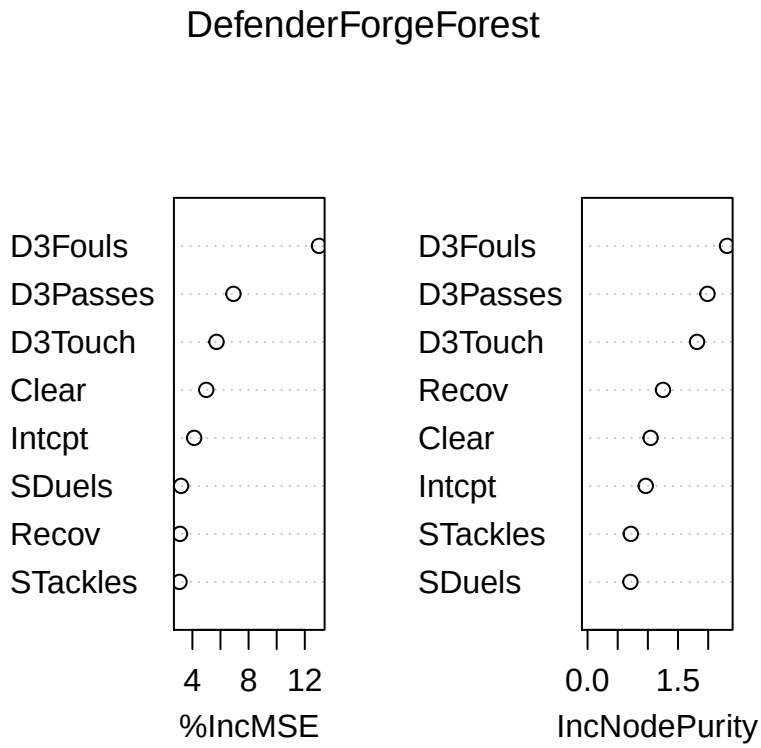
We will finally look at defenders, where we will have access to seven positions within the dataset:

- Central Defender
- Left Centre Back
- Right Centre Back
- Left Back
- Right Back
- Left Wing Back
- Right Wing Back

The statistics (and abbreviations) that will be used for this updated dataset are:

- Successful Duels Per Minute -> SDuels
- Successful Tackles Per Minute -> STackles
- Interceptions Per Minute -> Intcpt
- Recoveries Per Minute -> Recov
- Passes Completed in the Defensive Third Per Minute -> D3Passes
- Touches in the Defensive Third Per Minute -> D3Touch
- Fouls in the Defensive Third Per Minute -> D3Fouls
- Total Clearances Per Minute -> Clear

In this section, we will focus on Forge FC and their head coach Bobby Smyrniotis.



Above all else, there is a large focus on fouls in the defensive third. Bobby Smyrniotis seems to be making defensive substitutions on the basis of not conceding fouls in the final third.

3 Further Work

We must take caution with a lot of the statements about coaching substitutional styles. While the models used were relatively good at explaining the variability of the dataset, further analysis should include much more data (within the same substitutional format). Using different variables when building similar models can hone in on statistics that a team would want to focus on when deciding upon a manager. Non-performance statistics like yellow cards could also be added to more general models.

An optimal dataset would contain the statistics of players at the moment of substitution. While we can get a good estimate from this dataset, tactical and game state changes can alter a player's statistics in short time. Game state itself is another possible variable that could be looked into. Head coaches may change

their substitutional strategy around when they are trailing compared to when they would like to preserve a lead. Formational differences might also unveil some important information. Specifying injury substitutions and regular substitutions will also add accuracy. Extending position categories can find more nuanced conclusions. For example, seperating the midfield category into attacking midfielders, wide midfielders, and defensive midfielders may allow for more position relevant statistics to be used. These are all possible extensions of this work, but they do rely on an improved dataset.

Another possible extension from this work is the assessment of substitutional value. There are now plenty of statistics like Valuing Actions by Estimating Probabilities (VAEP) and Goals Added (G+) that assess a lot of the major components of the sport. Those statistics can be used to assess the difference in value a substitute makes compared to the player subbed off.

This paper also relys on the notion that managers have agency to sub players off based off of their performance. Substitutions can be used for formational purposes, rather than performance ones. With this, managers may not be focused on who is performing badly, but where a change needs to be made to turn the tide of a game. Saying that coaches are making decisions purely on how player may be hyperbolic, but it may reveal some unconcious biases that managers may have.

4 Discussion

While a large focus has been placed on how players are performing in soccer analytics, there has not been a huge push to assess and analyze managerial behaviors. Advancing metrics on the coaching side can help identify managers who can read the game well and make decisive decisions. Managers who have been unlucky in their match outcomes can be identified and recruited for less than they should actually be worth. Teams who can appropriately use coaching statistics can set themselves up for sustainable success in the future.

5 References

Decroos, Tom, et. all, "Actions Speak Louder Than Goals: Valuing Player Actions in Soccer", July 2019, https://dl.acm.org/doi/10.1145/3292500.3330758.

Muller, John, "Goals Added: Introducing a New Way to Measure Soccer", May 4 2020, https://www.americansocceranalysis.com/home/2020/4/22/37ucr0d5urxxtryn2cfhzormdziphq?rq=goals%20added.

Myers, Bret, "A Proposed Decision Rule for the Timing of Soccer Substitutions", 2011, https://cafefutebol.files.wordpress.com/2013/12/substitution_timing.pdf

Stats Perform, "Canadian Premier League Centre Circle Data", 2019, https://canpl.ca/centre-circle-data