Supervised Learning Approach for Improving Data Quality in Sports

Changxuan Liu, Kiron Shastry, Matthew A. Lanham
Purdue University, Department of Management, 403 W. State Street, West Lafayette, IN 47907
Athlyte, Inc., 800 22nd Avenue, Tuscaloosa, AL, 35401
liu2371@purdue.edu; kiron@athlytesports.com; lanhamm@purdue.edu;

Abstract

Sports analytics is increasingly using detailed play-by-play data to generate insights. One such example is the use of substitution information in basketball play by play data to develop analytics such as the basketball APM. However, the errors in the substitution data influence the insights developed through this analysis. The goal of this study is to analyze the structure and causes of the errors in substitution information and to develop methods for the correction of these errors in NCAA basketball game data. In the paper, a categorization of the errors is provided. In addition, details of the investigation done to find and correct the substitution data errors using advanced analytic techniques are presented.

Key words: Sports analytics, machine learning, predictive analytics, data quality, supervised learning, MongoDB, Python, and R

Introduction

Sports analytics has gained significant research interest in the past few years. The motivation of this study comes from the analytics around play by play data. Such research requires accurate substitution information in play-by-play data. Play-by-play data consists of play entries recording actions and incidents occurring during a game. For example, a typical play entry can record the game, time, play type, result from the play, the player(s) associated with play and the score of the two teams at the end of the play. Analytics can be performed by tracking the play-by-play data to generate insights of a team, a player, or a type of action like fouls and goals. For example, the Adjusted Plus/Minus (APM) model (Omidiran 2011) is an important tool for player's contribution analysis. However, such analysis requires high quality and comprehensiveness of the substitution records in the play-by-play dataset. Errors in the dataset introduces noise irreducible by improving modeling techniques. Errors like missing records and mistakes can be common and random from game to game and from second to second and thus cannot be easily removed. Wu (2017) attempted to improve the quality of play-to-play data in basketball, and dealt with errors in player substitutions by using artificial intelligence techniques. The objective is to improve the quality of substitution records by designing a predictive model to detect and correct the substitution errors found in play by play data.

The remaining sections are organized as follows. First, literature is reviewed to understand the state of research in this context. Next, the structure of the play-by-play data used to build and validate the model is presented. This includes categorizing the errors, investigating the occurrence of the errors, proposing a supervised model to identify the errors, and attempting a graphic approach to investigate substitution patterns to look for corrective replacements for the errors. Next, the results are presented after each section of the methodology. This paper is concluded with a reflection on the limitations and directions for future study.

Literature Review

Wu (2017)'s approach was focused on deciding whether substitution data is accurate or not by using the play-by-play data records occurring 'near' the examined substitution record. In the paper, substitution errors were categorized into four types:

> Error Type 1: games that have no player substitutions recorded at all.
> Error Type 2: game data with play by play records recording an unequal number of players substituted in when compared to the players substituted out.
> Error Type 3: player substitution patterns not alternating between in and out.
> Error Type 4: recording that a substitution occurred but not capturing the name of the player(s) being substituted.

Wu designed an agent which removed a substitution record or imputing a substitution that it believed should have been recorded when substitutions were allowed during the game. The agent was initialized with a set of players current on the basketball court. For each record substitution the agent assigned a confidence score based on contextual evidence and those records which didn't pass the classification threshold were removed. When a correct substitution record was encountered, the agent was updated. An important part of this agent is a binary classifier on whether a recorded substitution was correct or not, to ensure 5 unique players on the court for each team. To train the classifier, Wu obtained supervision for the model by manually recording the players using video footage of 5 games. One drawback of this approach is the small size of the training dataset because manually watching video footages were time-consuming. The classifier was trained using logistic regression. In the model proposed in this study, more supervised learning algorithms will be tested. After that, Wu calculated an activity heuristic score for each player to infer the record to be imputed once a missing record was detected by the agent. The player with the highest activity heuristic score was added to the agent's tracking of players. Wu's methodology relied on comparison between the play-by-play data and the game footage. This paper sought to address the substitution errors using the information in the play-by-play dataset.

The paper by Lucey et al (2014) used play-by-play data to predict chances of open shots in a basketball game. It was shown that the number of defensive role-swaps is predictive of getting an open-shot and this measure can be used to measure the defensive effectiveness of a team. The study didn't involve predictive modeling, but used a series of T-tests to identify significant factors. In this paper, a similar approach is used to identify the factors to predict and correct substitution errors.

Data

In this paper, play-by-play data from 50 basketball games of two NCAA teams is utilized to perform the analysis. This data includes data from the University of Alabama (UA) women's basketball and Texas A&M University (TAMU) men's basketball games. The data analyzed is shown in Table 1.

Team Name	Sport	Season	Number of Games	Periods of a game
TAMU	Men’s Basketball	2017-2018	35	2 20-minute periods
UA	Women’s Basketball	2018-2019	15	4 10-minute periods

Table 1 Data coverage

For each game, the play-by-play data includes the following fields:

- team:	string;	the team that the play was associated with
- vh:	binary;	whether the associated team is playing as home (h) or visitor (v)
- time:	time;	the clock time when the play occurred (e.g. 15:30)
- checkname:	string;	name of the player associated with the play
- uni:	string;	uniform number of the player
- vscore:	numeric;	the visitor team score by the time play occurred
- hscore:	numeric;	the home team score by the time play occurred
- action:	string;	the action of the play
- type:	string;	the type of the action
- qualifier:	string;	the characteristic of a goal if the action is a goal (e.g. fast ball)

A more detailed explanation of the action and type fields is given in Appendix 1. In addition, the following NCAA rules and common practices will influence substitution decisions:

(1) There are no limits to the number of substitutions that a team can make;

(2) Substitutions can be made when the ball is dead and the clock is dead. Apart from those marked as DEADBALL after a rebound, a ball is also dead and substitutions are permitted at period breaks, after a foul, and at timeouts.

(3) A player is fouled out after 5 fouls. The player can’t play for the rest of the game . Substitution has to be made for a fouled-out player.

(4) Typically, the start players are substituted out after 8 to 10 minutes of playing. Start players typically play for 30 to 40 minutes of a game cumulatively, while substitute players play for around 20 minutes.

Methodology and Analysis

1. Categorization of substitution errors

1.1 Data model

Each substitution consists of two play entries, hereafter referred to as a substitution pair, with a substitution-out record, which describes the player who is being substituted out and is followed by a substitution-in record. In general, the substitution errors can be classified into two types, fully missing error and partially missing error. A fully missing error occurs when the dataset is missing the entire substitution pair. In other words, the substitution was not recorded. On the other hand, a partially missing error has either the substitution-in or the substitution-out record missing. The partially missing errors can be easily detected by searching play-by-play data for unmatched substitution pairs. However, correcting these errors can be difficult because the play-by-play data only records a limited range of actions and the number of substitutions is unrestricted in basketball. Here is an example of a difficult situation. When a player is substituted in but his substitution-in record was missing. During the time he was on the court, he didn’t make any actions that would result in a mention in the play-by-play data. Then this player was substituted out and a data error occurred again at this point because his substitution-out was also not captured. As a result, the information that this player had played was totally missing and imputing the missing substitution-in record was challenging.

To better classify substitution errors, the play-by-play dataset of a game is exhaustively partitioned into substitution windows and game times. A ‘substitution window’ is defined as a sub-set of play-by-play data with a time when substitutions were allowed. According to the rules, substitution windows start at the occurrence of a foul, timeout, period break, and dead balls. When such actions happen the game clock is stopped so the play records in the same substitution window have the same timestamps. The time when the game was continuously undergoing between two consecutive substitution windows is referred to as a ‘Game Time’. When a substitution window ends, the game continued and a ‘Game Time’ starts. The play records within the same game time will have different timestamps. In this way, game times and or substitution windows are exclusive and exhaustive partitions for a game. Substitution errors can only exist in substitution windows. Figure 1 looked at a home game of UA played on January 3, 2019, and took a few play-by-play records as an example to illustrate how the play-by-play dataset was partitioned into substitution windows and game times.

Play-by-play data records					Partitions
clock	checkname	team	type	action	
…	…	…	…	…	… **Game Time X-1**
7:36	JOHNSON_CIERRA	UA	others	FOUL	Foul. Clock stopped. Substitution Window, X, began. **Substitution Window X** Substitutions were allowed and made during a window.
7:36	JOHNSON_ARIEL	UF	IN	SUB	
7:36	WILLIAMS_ZADA	UF	IN	SUB	
7:36	ROBINSON_PAIGE	UF	OUT	SUB	
7:36	SMITH_KIARA	UF	OUT	SUB	
7:26	JOHNSON_ARIEL	UF	3PTR	GOOD	The substitution window ended. The game went back on. **Game Time X**
7:13	JOHNSON_CIERRA	UA	JUMPER	MISS	
7:13	WILLIAMS_ZADA	UF	DEF	REBOUND	
6:48	WASHINGTON_DELICIA	UF	JUMPER	GOOD	
6:29	ABRAMS_MEGAN	UA	others	TURNOVER	
6:27	WASHINGTON_DELICIA	UF	others	STEAL	
6:23	JOHNSON_ARIEL	UF	3PTR	MISS	
6:23	COPELAND_ARIYAH	UA	DEF	REBOUND	
6:13	JOHNSON_CIERRA	UA	3PTR	MISS	
6:13	WILLIAMS_ZADA	UF	DEF	REBOUND	
5:55	NAKKASOGLU_FUNDA	UF	others	TURNOVER	
5:39	JOHNSON_CIERRA	UA	JUMPER	MISS	
5:39	WASHINGTON_DELICIA	UF	DEF	REBOUND	
5:30	WILLIAMS_ZADA	UF	JUMPER	GOOD	
5:30	WASHINGTON_DELICIA	UF	others	ASSIST	
5:16	WALKER_JASMINE	UA	JUMPER	MISS	
5:16	WASHINGTON_DELICIA	UF	DEF	REBOUND	
5:07	WILLIAMS_ZADA	UF	JUMPER	GOOD	
5:07	WASHINGTON_DELICIA	UF	others	ASSIST	
4:43	WALKER_JASMINE	UA	JUMPER	MISS	
4:43	NAKKASOGLU_FUNDA	UF	DEF	REBOUND	
4:18	NAKKASOGLU_FUNDA	UF	JUMPER	MISS	
4:18	WILLIAMS_ZADA	UF	OFF	REBOUND	
4:16	WILLIAMS_ZADA	UF	JUMPER	MISS	
4:16	COPELAND_ARIYAH	UA	DEF	REBOUND	
4:08	JOHNSON_CIERRA	UA	others	TURNOVER	
4:08	JOHNSON_ARIEL	UF	others	STEAL	
4:08	TEAM	UF	MEDIA	TIMEOUT	Timeout. Clock stopped. Another Substitution Window, X+1, began. **Substitution Window X+1**
4:08	SMITH_KIARA	UF	IN	SUB	
4:08	ROBINSON_PAIGE	UF	IN	SUB	
4:08	WILLIAMS_ZADA	UF	OUT	SUB	
4:08	NAKKASOGLU_FUNDA	UF	OUT	SUB	
4:08	KNIGHT_ASHLEY	UA	IN	SUB	

4:08	COPELAND_ARIYAH	UA	OUT	SUB		Substitutions were allowed and made during a window.
…	…	…	…	…	… **Game Time X+1**	

Figure 1 Partition the play records into substitution windows and game times

If Window X is taken as a substitution window occurred in the middle of the game, there are 3 possible conditions of a substitution pair at Window X.

(1) Both the in record and the out record are complete, so the substitution pair has no errors.
(2) Either the in or the out record is missing. This kind of error was relatively easier to detect. In addition, these kinds of errors can be further categorized into 'too-few' errors, where only the substitution-in record was missing and the opposite, 'too-many' errors.
(3) The entire substitution pair was not recorded.

Table 2 gives verbal definitions to the error categories and illustrates the categorization taking into consideration the substitution windows and game time immediately before and after Window X, referred to as Window X-1, Window X+1, Game Time X-1, and Game Time X. In this scenario, Player A was supposed to be substituted out and Player B was going to take A's place.

Error Category			Definition
No errors			Both in and out record of a substitution pair exist.
Detectable errors	Too few: The record exists for someone substituted out but no record for a new player being substituted in, resulting in less than 5 players in the subsequent game time for the associated team	Correctable-easy	Plays associated with the player of the missing substitution record can be found in the next game time or the next substitution window.
		Uncorrectable	Plays associated with the player of the missing substitution record cannot be found in any other game times or substitution windows.
		Correctable-hard	Plays associated with the player whose substitution record is missing don't exist in the immediate game time or substitution window, but can be found in later game times or substitution windows.
	Too many: The record exists for someone substituted in but no record for a new player being substituted out, resulting in more than 5 players in the subsequent game time for the associated team	Correctable-easy	Plays associated with the player of the missing substitution record can be found in the previous game time or the previous substitution window.
		Uncorrectable	Plays associated with the player of the missing substitution record cannot be found in any other game times or substitution windows.
		Correctable-hard	Plays associated with the player whose substitution record is missing don't exist in the immediate antecedent game time or substitution window, but can be found in later game times or substitution windows.
Substitution not recorded			Both records of a substitution pairs are missing. These errors can also be further categorized in similar patterns to the detectable errors, depending on whether there are play records

	associated to the players in other game times and substitution windows.

Table 2-a Definition for error categories

The following table is an illustration of the above categorization of substitution errors. '+/-' represents the existence of missing of play records. For example, 'A out +' means the record for player A being substituted out exists. 'B in –' means the record for player B being substituted out is missing. 'Play(A) +' means a play record associated with player A of any type exists in a game time.

Timeline of scenario Error Category		Window X-1	Game Time X-1	Window X		Game time X	Window X+1
No Errors		/	/	A out +	B in +	/	/
Detectable-Too few	Correctable-easy	/	/	A out +	B in -	Play(B) +	/
		/	/	A out +	B in -	Play(B) -	B out +
	Uncorrectable	/	/	A out +	B in -	Play(B) -	B out -
	Correctable-hard	/	/	A out +	B in -	Play(B) -	B continued playing
Detectable-too many	Correctable-easy	/	Play(A) +	A out -	B in +	/	/
		A in +	Play(A) -	A out -	B in +	/	/
	Uncorrectable	A in -	Play(A) -	A out -	B in +	/	/
	Correctable-hard	A continued playing	Play(A) -	A out -	B in +	/	/
Substitution not recorded		/	/	A out -	B in -	/	/

Table 2-b An illustration of different kinds of errors at Window X

The investigation of this study primarily focused on the correctable-easy errors (colored in green). These two types of errors will be jointly referred to as 'targeted errors' for the remainder of this paper. Other types of errors (colored in red) are beyond the scope of the analysis, though the supervised learning model proposed in the next section may provide some insight for resolving other types of errors.

Additionally, the following assumptions are made to ensure the feasibility of the analysis. First, the play-by-play data of the two major teams, TAMU and UA, and their home games have perfect data without any substitution errors. Second, overtime periods are not considered for this analysis because overtime is rarely observed in the data set available. Third, it is assumed that no substitution errors happened in two consecutive substitution windows. Fourth, the quality of the original data set is good enough so that only one error can occur at each substitution window.

1.2 Investigation result on targeted categories

The correctable-easy errors are found by comparing the number of substitution-in and substitution-out for each team, at each substitution window. Out of the 24,547 play records for all 50 games, there are 82 targeted errors found, about 0.334% of the entire dataset. Out of the total 2580 substitution windows in the entire dataset, 74 of them contain targeted errors, reaching a percentage of 2.87%.

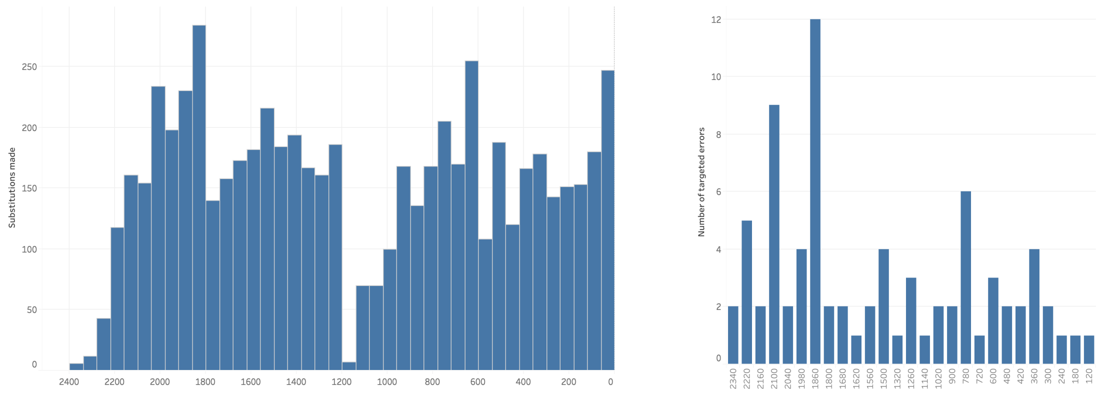


Figure 2 Time distribution of substitutions and substitution errors

Figure 2 shows the histogram of the time when substitutions were made, and the time when substitution errors occurred. In general, substitutions were made with similar patterns in the first half and the second half of the game. Substitution errors are more likely to occur in the first half of the game.

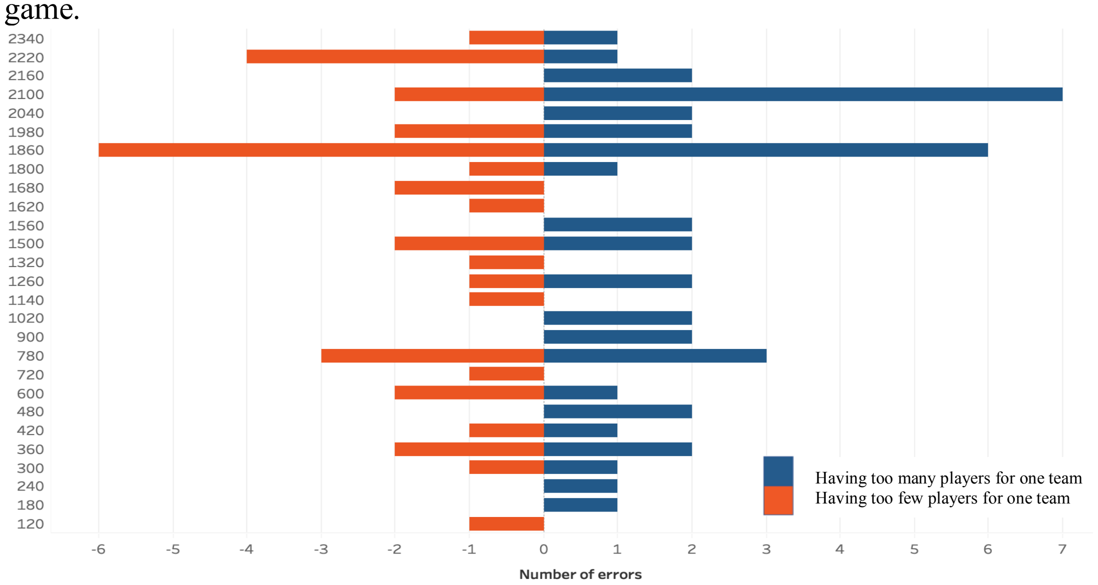


Figure 3 Time distribution of the two types of targeted errors

Figure 3 displays the correctable errors that result in too many players and too few players separately. Both types of errors are more likely to occur in earlier parts of the game. The substitutions at game breaks are more likely to be accurately recorded.

2. A supervised predictive approach

Wu's model (Wu, 2019) aimed to directly determine the correctness of a record. However, identifying all possible situations is challenging and complex. Also, the low error occurrence rate results in a low number of samples to train the classifier. This study seeks to avoid directly classifying a record as an error or a correct one. Instead, attempts are made to propose a supervised learning approach to predict whether a player should be substituted out at a substitution window, given the player's performance, the time he has played in the game, and the situation in the game at that time. Assuming the play-by-play data of TAMU and UA with the home games are free of any errors, the models are trained on a perfect play-by-play dataset first and then can be tested on an imperfect dataset created by randomly removing some records to introduce errors. However, whether substitution errors occur does not necessarily affect whether a player should be substituted out. In other words, all the substitution-out records that already existed in the play-by-play data set are valid.

2.1 Feature Engineering

Recall that the model is a supervised binary classifier with y representing whether or not a player should be substituted out at a substitution window. A player's performance is evaluated according to the progress of the game instead of the entire game in order to imitate the information available to the coach during the game. The features used for training the model falls into three categories:

(1) Game situation so far

This category included the situation on a team level, such as whether the team is taking a lead, how long has the team been leading, and the size of the lead, etc. The progress of the game including how much time was left and which period it was also falls into this category.

(2) Player cumulated performance

This is a cumulative record for the actions that a player has performed for his or her time in game up to a moment. Some extraordinary actions such as dunks and 3-point goals has higher power to tell the good performance, though may happen less frequently. The number of actions is taken as a percentage of the same kind of actions made by the entire team to reduce bias. These features are also expected to partly capture the coaches thinking and forecasting for the remaining part of the game when making a substitution decision.

(3) Player temporary performance

Considering that a player may have been substituted out and in multiple times, this category captures a player's statistics for the current period of time that he has been playing in game. In other words, this category includes the performance since the last time the player was substituted in. The stats are also taken as a percentage of the same kind of actions for the entire team during the same time frame.

2.2 Modeling

Models are trained using CARET package in R. Z-score standardization is applied to numeric variables. The categorical variables are one-hot encoded into dummies. Linear combos are removed. However, variables with near-zero-variance are kept due to many columns which are by definition small in magnitude. The dataset is partitioned into train and test set with ratio of 80% to 20%. 5-fold cross validation is used to adjust the tuning parameters to search for the best model. The primary model performance measure used to compare models is ROC. The secondary performance measure is Matthew's Correlation Coefficient. The models tested include generalized linear models, C5.0 tree, adaBoost, XGBoost, svm, and artificial neural network. The models are checked for overfits using validation set approach. Candidate models are those with the less than 10% difference in performance between train and test.

To further balance the bias and variance, the performing models are assembled using propensity averaging. The weights assigned to each model is based on the model's balanced accuracy because the goal is to achieve the highest performance on deciding the substitution-outs.

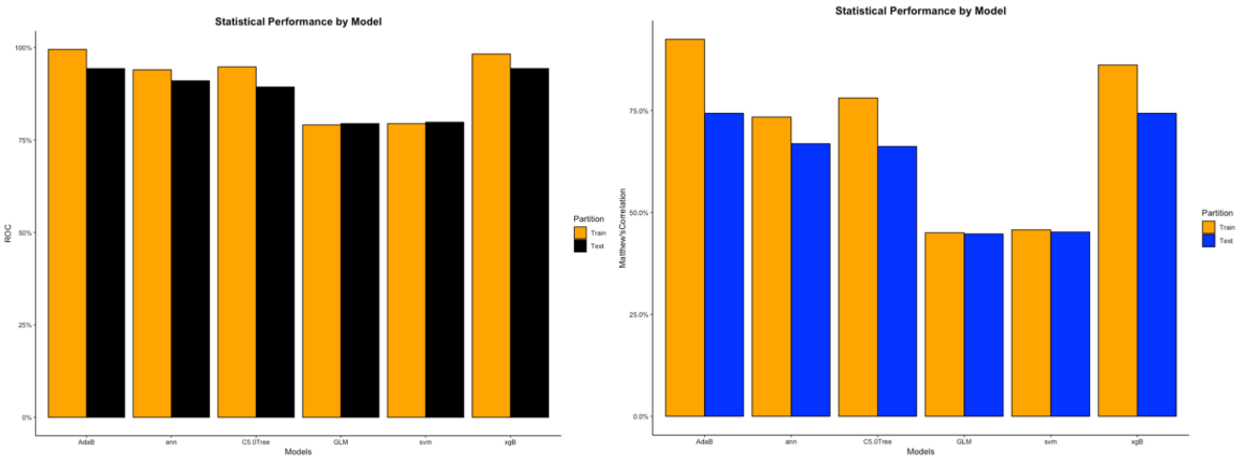
2.3 Supervised modeling results

The binary classifier is trained using the home games for UA and TAMU to predict whether a player was substituted out. The dataset used for model training and validation has 12,993 rows and 49 predictors with dummy variables included. A detailed list of features is displayed in Table 3.

Feature category	Variable	Data type	Explanation
Response variable	y	Factor, binary	Response variable, whether the player was substituted out
Team condition so far	timeleft	Numeric	How much time (seconds) was left till the end of the game, assuming no overtime
	sport	Factor, binary	Men's or Women's basketball
	lead	Factor, binary	Whether the team of the player is in a leading position
	lead_duration	Numeric	How many seconds has the team been taking a lead
	lead_size	Numeric	The difference in scores, negative if falling behind
	played_games	Numeric	How many games has the team played in this season
	rival	Factor	Which opponent is the team playing against
	score	Numeric	Total score of the team
	homegame	Factor, binary	Whether the team is playing at home
Player cumulated performance	perc_score_sofar	Numeric	Scores by player / score by team
	turnover_sofar	Numeric	Turnovers by player
	steal_sofar	Numeric	Steals by player
	good_rate_sofar	Numeric	A player's scoring effectiveness, total goals / total attempts
	score_per_second_sofar	Numeric	A player's scoring efficiency, total scores / time played
	fouls	Numeric	Total fouls made (Risk of being penalized out)
	starter	Factor, binary	Whether a player is on the starting roster
	dunks	Numeric	Number of dunks successful
	blocks	Numeric	Number of blocks successful
	3pt_good	Numeric	Number of 3-point throws successful
Player temporary performance	time played	Numeric	How long has the player been playing without rest (stamina)
	score_temp	Numeric	Scores by player since the player most recently became active (the start of the game for starters)
	turnover_temp	Numeric	Turnovers by player since the player most recently became active (the start of the game for starters)
	steal_temp	Numeric	Steals by player since the player most recently became active (the start of the game for starters)
	good_rate_temp	Numeric	A player's scoring effectiveness, total goals / total attempts since the player most recently became active

Table 3 List of features used for modeling

According to Figure 4, when ROC is used as the performance measure, the models yielded similar performance on the test data set. None of the models are overfitting. Thus, Matthew's Correlation Coefficient (MCC) is used to further separate the model's performance. In terms of MCC, AdaBoost is overfitting, and the best performing model is XGBoost. According to Table 4, The propensity averaged ensemble model failed to outperform XGBoost.



ROC Matthew's Correlation Coefficient

Figure 4 Model performance comparison

Models	ROC_train	ROC_test	MCC_train	MCC_test
GLM	0.7931	0.7949	0.4578	0.4506
C5.0Tree	0.9518	0.8946	0.7890	0.6822
AdaB	0.9948	0.9491	0.9289	0.7637
xgB	0.9949	0.9667	0.9278	0.8047
svm	0.7965	0.7979	0.4604	0.4567
ann	0.9369	0.9095	0.7138	0.6550
Ensembled	0.9873	0.9528	0.9035	0.7623

Table 4 Model performance comparison

500 substitution records are taken out of a perfect dataset to create an imperfect dataset against which the model is tested. In the end, the model is able to successfully locate 98% of the records deliberately removed.

3. A graphical approach to substitution habits

3.1 Analysis method

Substitution habits are defined as players who usually replace each other. Substitution habits may exist between players of the same position or similar skills. Discovering substitution habits is not the focus of this study, but we tried a graphical approach to substitution patterns. Different from the supervised model above which focused on who should be substituted out, investigating on substitution habits focused on who should be imputed when the substitution-in record was missing.

Substitution habits are searched by comparing the status of a player, whether playing on the floor or resting on the bench. If play records exists for a player during a game time, a player is marked as in game (1). Otherwise, the player is marked as out of game (0). All players are tracked throughout a game for being in game or out of game. The status of players is plotted into line charts together against time. The intuition behind this approach is that, if two players are always in game together, they are unlikely to be substitutes for each other. Otherwise, if a player is always out of game while the other is in game, they are likely to belong to the same substitution pair. In other words, if the line chart of two players 'compensate' each other, they are likely to be substituting players.

3.2 Results on substitution habits

The game of UA on 1/3/2019 is taken as an example. The status of players throughout the game is plotted as Figure 5. It is obvious that whenever BerryTaylor (top line) was in game, Abrahams_Megan (second top line) was always out of game. Their status was always compensating each other. Hence, it is very likely that these two players are substitutes for each other. Therefore, when there is substitution errors associated with Berry_Taylor, Abrahams_Megan should be the primary choice for correction and vice versa. However, this graphic representation is limited to substitution habits involving only 2 players. And it doesn't look at all games for UA at the same time. The purpose of including it is to provide some intuition for future investigations.

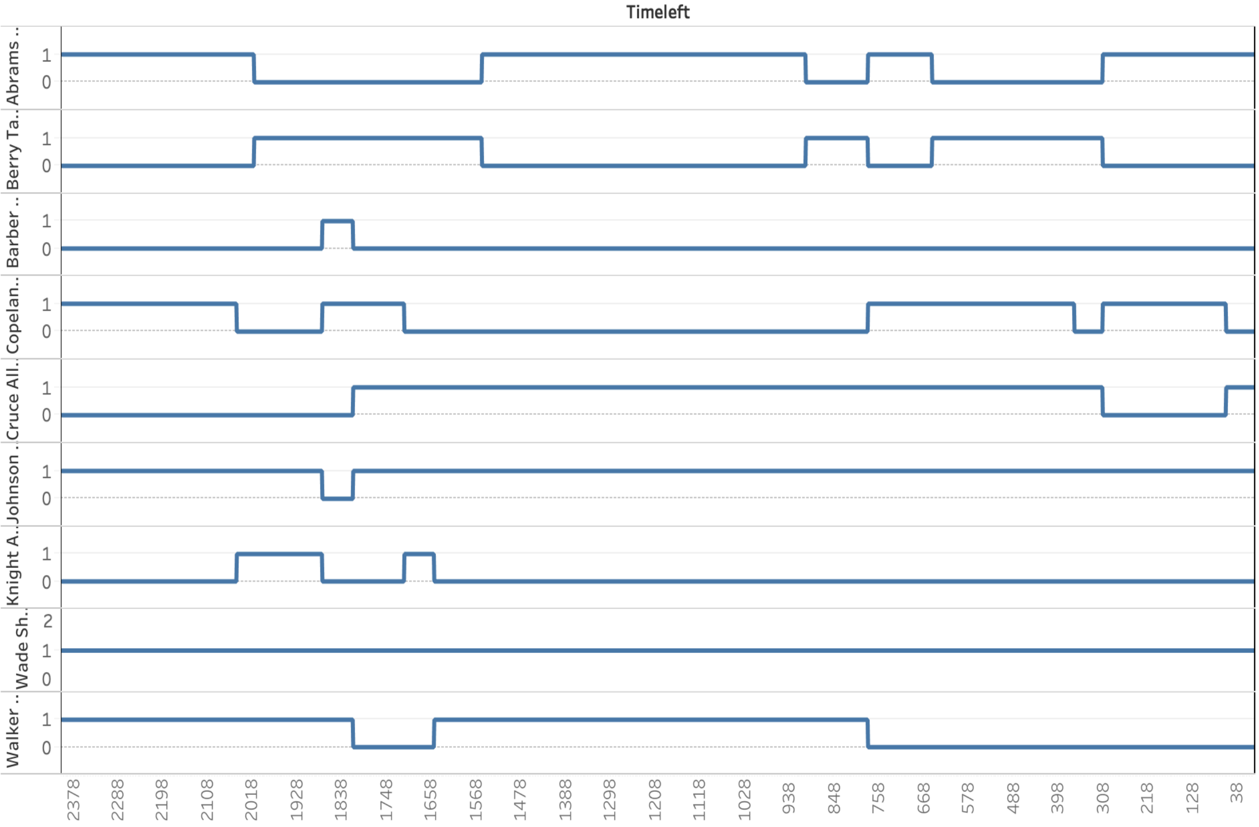


Figure 5 Graphic representation of player status (1: in game, 0: out of game)

Conclusion

Sports analytics based on play-by-play data largely relies on high data quality. This study addresses the errors in substitution records by providing a detailed categorization of possible types of substitution errors, and proposes a binary classifier to detect and correct two types of errors. Among the supervised learning algorithms, XGBoost achieves superior predicting performance. The proposed model attempts to detect substitution errors by comparing the available dataset with the prediction result, instead of directly assessing whether a record is correct or mistaken. The final model is able to predict the substitutions with ROC about 95%. Plus, a graphical approach to look for substitution habits is demonstrated to identify pairs of players usually substituting each other. The purpose of this is just to provide some preliminary insights in how the errors can corrected. There are four major limitations with the methodology of this study.

(1) The assumption that the data set is of reasonable quality so as to eliminate some of the more complex errors has not been validated.
(2) The proposed supervised learning model can only work on errors with substitution-out records. Its effect on the errors with substitution-in records was limited.
(3) The play-by-play data set in this study is not comprehensive, recording only a certain range of actions.
(4) The substitution habits can involve more than 2 players, and thus become much more complex.

Future study is advised as follows:

(1) Extending Wu's methodology by watching more game videos for better modeling;
(2) Investigation of fully-missing errors and partially-missing-puzzle/hard errors;
(3) Utilizing deep learning models to directly analyze the game videos to identify the frames of a substitution;
(4) Investigation of substitution habits to come up with the player name for a missing substitution record.

References

Lucey, P., Bialkowski, A., Carr, P., Yue, Y., & Matthews, I. (2014). How to get an open shot: Analyzing team movement in basketball using tracking data. In *Proceedings of the 8th annual MIT SLOAN sports analytics conference*.

Omidiran, D. (2011). A new look at adjusted plus/minus for basketball analysis. In MIT Sloan Sports Analytics Conference [online].

Wu, S., & Swartz, T. B. (2017). Using AI to correct play-by-play substitution errors. MIT Sloan Sports Analytics Conference.

Appendix 1 Explanation of Actions and Types in play-by-play data

Action	Type	Explanation	
SUB	IN	Substitution a player in	
	OUT	Substitution a player out	
TURNOVER		A team loses possession of the ball to the opposing team before a player takes a shot at their team's basket.	
STEAL		A defensive player legally causes a turnover by his positive, aggressive action(s).	
ASSIST		Attributed to a player who passes the ball to a teammate in a way that leads to a score by field goal	
FOUL		An infraction of the rules more serious than a violation	
BLOCK		A defensive player legally deflects a field goal attempt from an offensive player to prevent a score	
REBOUND	OFF	Colloquially referred to as a board, is a statistic awarded to a player who retrieves the ball after a missed field goal or free throw.	Ball goes to offense team after rebound
	DEF		Ball goes to defense team after rebound
	DEADBALL		Ball goes dead
GOOD/MISS	3 PTR	3-point throw	
	JUMPER	Jump shot, score a **basket** by leaping straight into the air.	
	LAYUP	A two-point shot attempt made by leaping from below, laying the ball up near the basket, and using one hand to bounce it off the backboard and into the basket.	
	FT	Free throw after a rival team foul, 1 point if goal	
	DUNK	Dunk	
	TIPIN	Touching a ball into the basket as it bounces off the basket or board after a missed shot.	
TIMEOUT	MEDIA	NCAA allows timeouts for electronic media	
	30 SEC	30-second timeouts	
	20 SEC	20-second timeouts	