What Tactics Should Coaches Prioritize in Soccer?

By Patrick Geraghty and Zidong Liu

Introduction

- Sports analytics is a relatively new domain area that is rapidly growing
- The growth of analytics within soccer specifically has been slower than that of other popular sports
- Challenges
  - The unpredictable nature of soccer and its varied play styles.
  - Resistance from traditional coaching methods.
- We will use performance related metrics that measure tactical principles to see what tactics coaches really should be prioritizing
  - Using data from 2022 and 2023 seasons of Major League Soccer

State of the Art

Frequency of r-pass moves in F.A. matches—negative-binomial test

r	Original data			Excluding 0-pass shots		
	Actual	*Expected*	*A − E*	*Actual*	*Expected*	*A − E*
0	10,580	10,542	+38	10,187	10,143	+44
1	6,923	7,975	−152	6,923	7,022	−99
2	3,611	3,542	+69	3,611	3,553	+58
3	1,592	1,572	+20	1,592	1,578	+14
4	608	653	−45	608	651	−43
5	280	260	+20	280	257	+23
6	107	101	+6	107	98	+9
7	33	38	−5	33	37	−4
8	9	14	−5	9	13	−4
9 and over	11	8	+3	11	8	+3
Total	23,754	23,805		23,361	23,360	
Mean	1·00			1·02		
Variance	1·495			1·496		
$P(\chi^2)$			>0·10			>0·20

Fitted by equating the mean to r/c and the variance to $[r(c+1)/c^2]$.

- The analysis of passing sequences.(1968)

Table 1. Passing performance in the 5 minutes before and after a goal is cored in relation to the overall half of the match where the goal was scored.

Sample / Performance indicator	Half of match goal was scored in	5 mins before goal	5 mins after goal	Friedman test
Scoring team				
Number of passes	23.2±5.2	22.5±8.8	21.5±11.1 ^	p < 0.001
%Successful passes	70.2±7.5	72.4±12.7 ^	67.3±14.7 ^&	p < 0.001
Conceding team				
Number of passes	22.9±4.3	19.3±8.4 ^	22.1±9.4 &	p < 0.001
%Successful passes	69.3±6.0	67.8±13.6	66.0±14.0	p = 0.118

- How passing patterns in the Premier League change before and after scoring.(2008)

State of the Art

```
=== Confusion Matrix ===

   a   b   c   d   <-- classified as
  24   0   9   0 |   a = FOR
   0  17   0   0 |   b = GK
   6   0  32   3 |   c = MID
   0   0   4  24 |   d = DEF
```

Figure 1- confusion matrix for position prediction

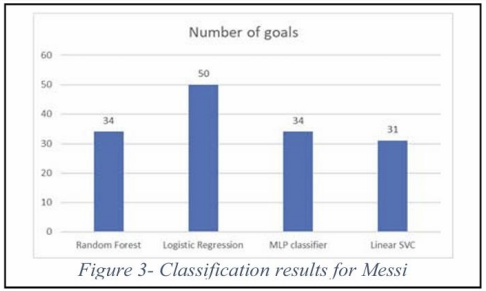
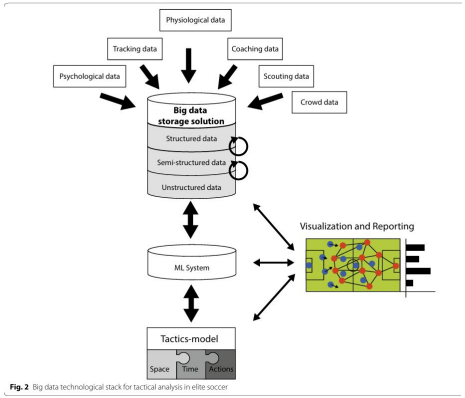
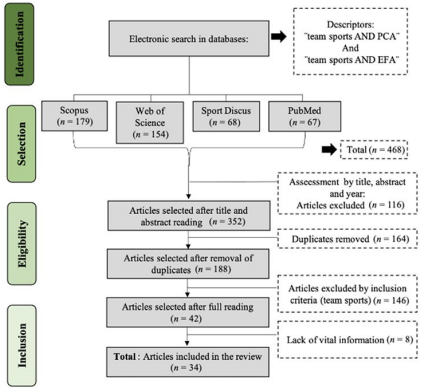


Figure 3- Classification results for Messi

- Use of machine learning to predict player positions and assess performance.(2019)





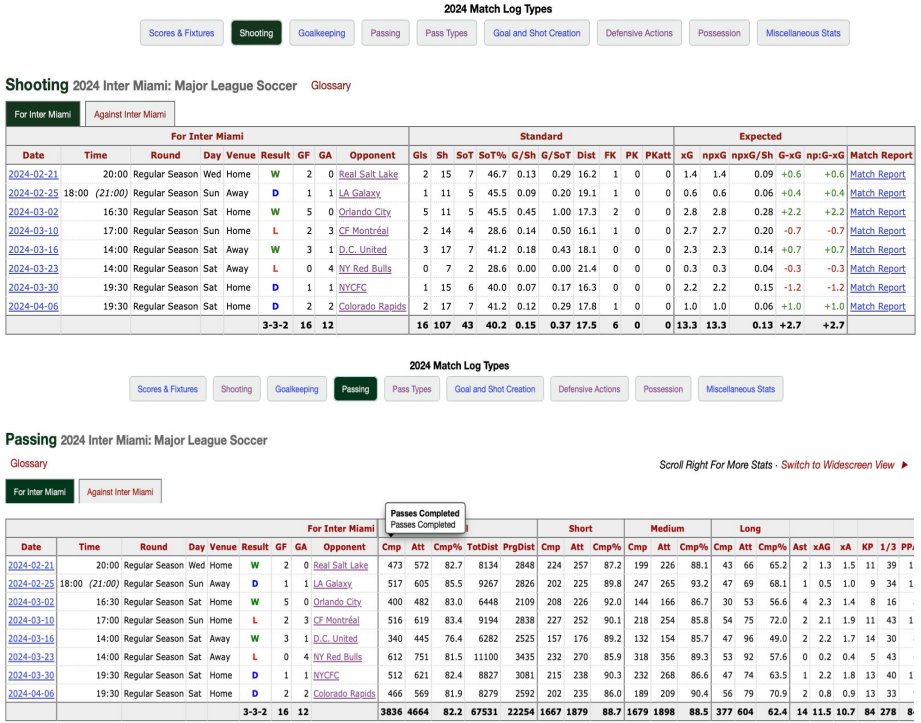
- How machine learning are transforming tactical analysis by identifying effective patterns and strategies.(2016)(2021)

Materials & Methods

- Dataset: We copy and pasted our data from individual MLS team match logs found at the FBRef website. All of this data is provided by Opta, or Stats Perform.
- Methods: We used logistic regression, linear discriminant analysis, and then a few tree based models (Bagging and Random Forests). We ran two different kinds of models, one with all our variables (only in tree based models) and another with metrics that measure tactical principles coaches could prioritize.
- Evaluation: As this is a classification problem, model accuracy and other confusion matrix derived metrics were used to evaluate our models. To evaluate and analyze our predictors we will use coefficients for LDA and logistic regression, and the mean decrease in accuracy for bagging and random forests

The Data

- Set 20% apart for testing, and performed cross validation on training data
- Seven different match logs: shooting, passing, pass types, shot/goal creation, defensive actions, possession, and miscellaneous
- Cut out all draws, redundant variables, and any irrelevant metrics
- Did some feature engineering to obtain common metrics used in soccer
- Created sublist of metrics measuring tactics coaches could prioritize



2024 Match Log Types

Scores & Fixtures | Shooting | Goalkeeping | Passing | Pass Types | Goal and Shot Creation | Defensive Actions | Possession | Miscellaneous Stats

Shooting 2024 Inter Miami: Major League Soccer Glossary

For Inter Miami | Against Inter Miami

For Inter Miami									Standard								Expected					
Date	Time	Round	Day	Venue	Result	GF	GA	Opponent	Gls	Sh	SoT	SoT%	G/Sh	G/SoT	Dist	FK	PK	PKatt	xG	npxG	npxG/Sh	G-xG	np:G-xG	Match Report
2024-02-21	20:00	Regular Season	Wed	Home	W	2	0	Real Salt Lake	2	15	7	46.7	0.13	0.29	16.2	1	0	0	1.4	1.4	0.09	+0.6	+0.6	Match Report
2024-02-25	18:00 *(21:00)*	Regular Season	Sun	Away	D	1	1	LA Galaxy	1	11	5	45.5	0.09	0.20	19.1	1	0	0	0.6	0.6	0.06	+0.4	+0.4	Match Report
2024-03-02	16:30	Regular Season	Sat	Home	W	5	0	Orlando City	5	11	5	45.5	0.45	1.00	17.3	2	0	0	2.8	2.8	0.28	+2.2	+2.2	Match Report
2024-03-10	17:00	Regular Season	Sun	Home	L	2	3	CF Montréal	2	14	4	28.6	0.14	0.50	16.1	1	0	0	2.7	2.7	0.20	-0.7	-0.7	Match Report
2024-03-16	14:00	Regular Season	Sat	Away	W	3	1	D.C. United	3	17	7	41.2	0.18	0.43	18.1	0	0	0	2.3	2.3	0.14	+0.7	+0.7	Match Report
2024-03-23	14:00	Regular Season	Sat	Away	L	0	4	NY Red Bulls	0	7	2	28.6	0.00	0.00	21.4	0	0	0	0.3	0.3	0.04	-0.3	-0.3	Match Report
2024-03-30	19:30	Regular Season	Sat	Home	D	1	1	NYCFC	1	15	6	40.0	0.07	0.17	16.3	0	0	0	2.2	2.2	0.15	-1.2	-1.2	Match Report
2024-04-06	19:30	Regular Season	Sat	Home	D	2	2	Colorado Rapids	2	17	7	41.2	0.12	0.29	17.8	1	0	0	1.0	1.0	0.06	+1.0	+1.0	Match Report
					3-3-2	16	12		16	107	43	40.2	0.15	0.37	17.5	6	0	0	13.3	13.3	0.13	+2.7	+2.7	

2024 Match Log Types

Scores & Fixtures | Shooting | Goalkeeping | Passing | Pass Types | Goal and Shot Creation | Defensive Actions | Possession | Miscellaneous Stats

Passing 2024 Inter Miami: Major League Soccer

Glossary

Scroll Right For More Stats · Switch to Widescreen View ▶

For Inter Miami | Against Inter Miami

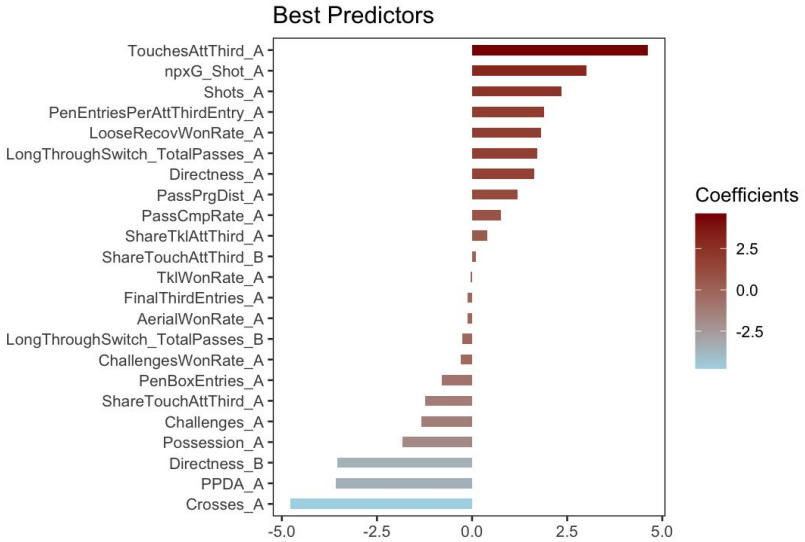
Passes Completed
Passes Completed

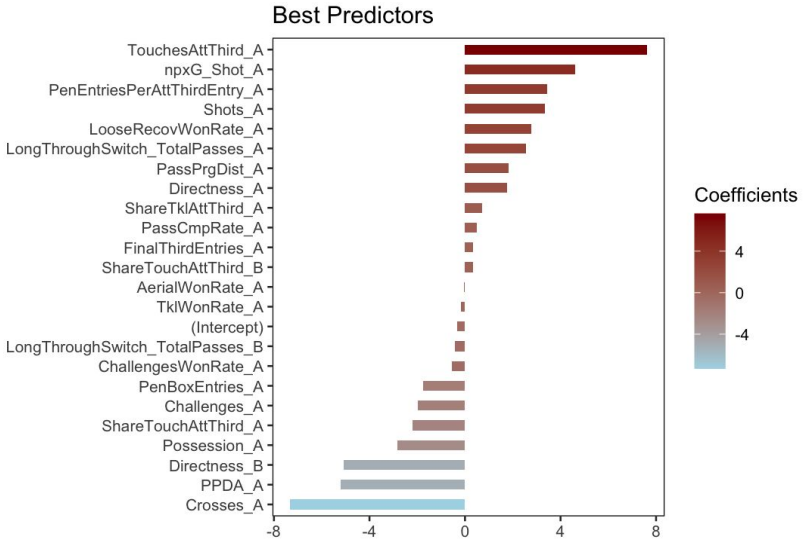
For Inter Miami														Short			Medium			Long								
Date	Time	Round	Day	Venue	Result	GF	GA	Opponent	Cmp	Att	Cmp%	TotDist	PrgDist	Cmp	Att	Cmp%	Cmp	Att	Cmp%	Cmp	Att	Cmp%	Ast	xAG	xA	KP	1/3	PP
2024-02-21	20:00	Regular Season	Wed	Home	W	2	0	Real Salt Lake	473	572	82.7	8134	2848	224	257	87.2	199	226	88.1	43	66	65.2	2	1.3	1.5	11	39	1
2024-02-25	18:00 *(21:00)*	Regular Season	Sun	Away	D	1	1	LA Galaxy	517	605	85.5	9267	2826	202	225	89.8	247	265	93.2	47	69	68.1	1	0.5	1.0	9	34	1
2024-03-02	16:30	Regular Season	Sat	Home	W	5	0	Orlando City	400	482	83.0	6448	2109	208	226	92.0	144	166	86.7	30	53	56.6	4	2.3	1.4	8	16	
2024-03-10	17:00	Regular Season	Sun	Home	L	2	3	CF Montréal	516	619	83.4	9194	2838	227	252	90.1	218	254	85.8	54	75	72.0	2	2.1	1.9	11	43	1
2024-03-16	14:00	Regular Season	Sat	Away	W	3	1	D.C. United	340	445	76.4	6282	2525	157	176	89.2	132	154	85.7	47	96	49.0	2	2.2	1.7	14	30	
2024-03-23	14:00	Regular Season	Sat	Away	L	0	4	NY Red Bulls	612	751	81.5	11100	3435	232	270	85.9	318	356	89.3	53	92	57.6	0	0.2	0.4	5	43	
2024-03-30	19:30	Regular Season	Sat	Home	D	1	1	NYCFC	512	621	82.4	8827	3081	215	238	90.3	232	268	86.6	47	74	63.5	1	2.2	1.8	13	40	1
2024-04-06	19:30	Regular Season	Sat	Home	D	2	2	Colorado Rapids	466	569	81.9	8279	2592	202	235	86.0	189	209	90.4	56	79	70.9	2	0.8	0.9	13	33	
					3-3-2	16	12		3836	4664	82.2	67531	22254	1667	1879	88.7	1679	1898	88.5	377	604	62.4	14	11.5	10.7	84	278	

Results

Model	Training Accuracy	Test Accuracy
LDA	75.89%	75.17%
Logistic Regression	74.99%	74.82%
Bagging (could prioritize)	72.43%	73.4%
Bagging (all variables)	83.42%	82.98%
Random Forest (could prioritize)	71.19%	73.05%
Random Forest (all variables)	83.6%	84.4%

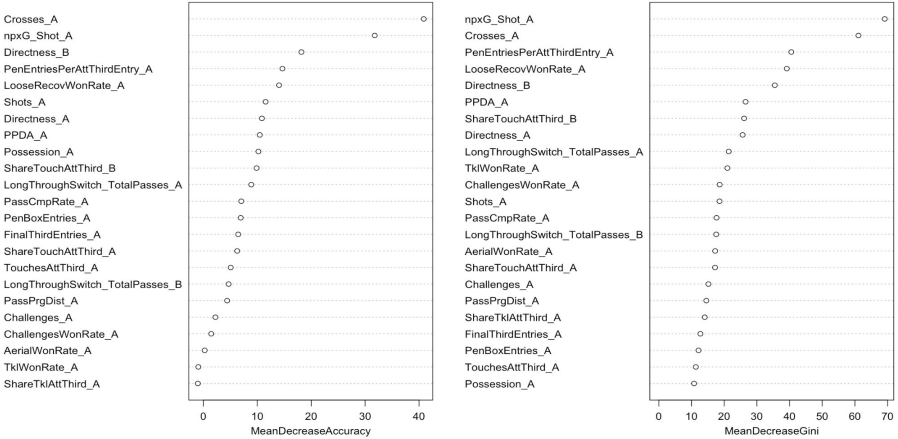
LDA and Logistic Regression



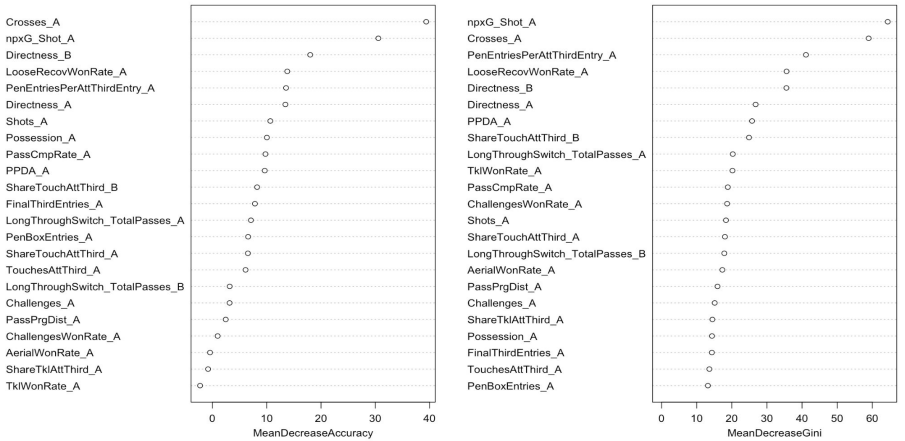


Bagging and Random Forests First Models

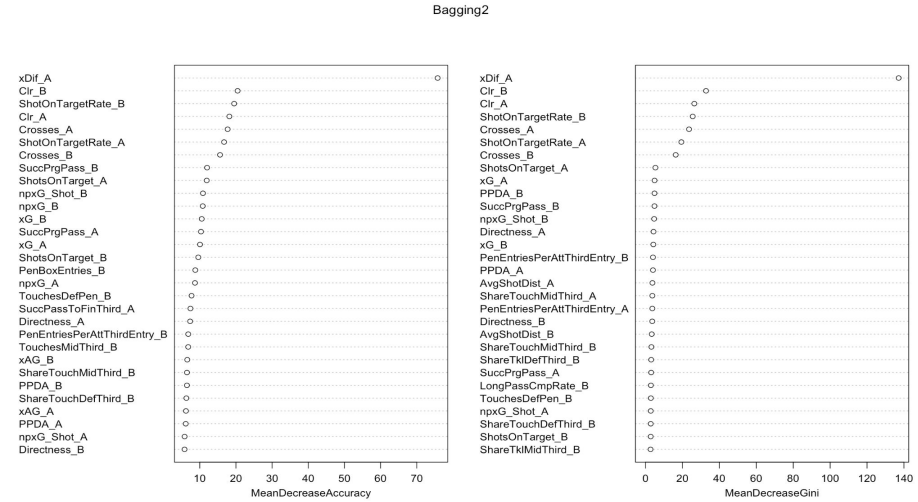
Bagging1

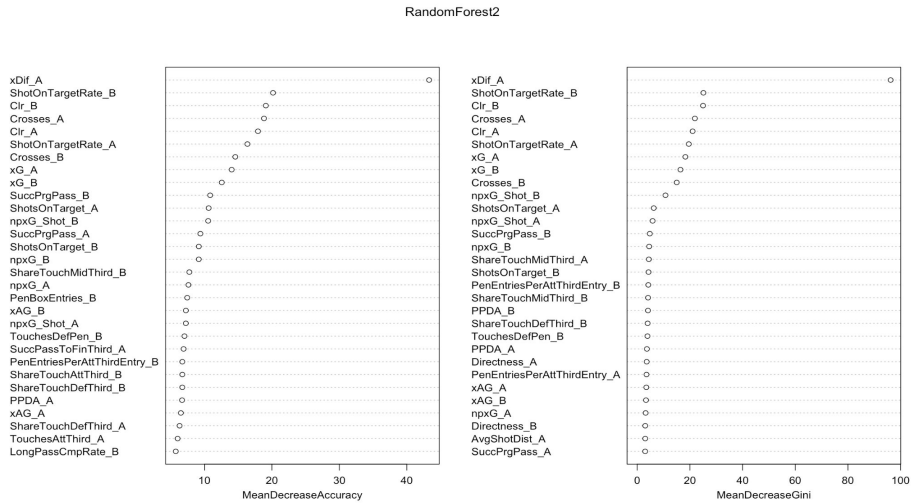


RandomForest1



Bagging and Random Forests Second Models





What Tactics Should Coaches Prioritize?

- First Model (including tactics coaches could prioritize)
  - Crosses (-), non-penalty xG per shot, and opponent directness (-) seem to be the most important when looking at the top five predictors by coefficient size (LDA/Log. Reg.) and mean decrease in accuracy (Bagging and Random Forests), as they appear each time
  - Passes per defensive action allowed, touches in the attacking third, 50/50 ball recovery rate, and the share of penalty box entries per attacking third entry are other metrics that appear to be relevant to our first model
- Second Model (all metrics)
  - One of the drawbacks with the second models is that we do not know the sign of the impact, being positive or negative, we only know overall impact as we used only used bagging and random forests for this second model
  - Given this, the five most “important” predictors based on mean decrease in accuracy are xDifference, clearances and opponent clearances (very surprisingly), crosses, and opponent shot on target rate

What Does This Actually Mean?

- First Model
  - Avoid “default” crosses
  - The importance of high quality shots
  - Preventing opposition progression
  - Defensive intensity, progression, winning 50/50 balls, and attacking third efficiency
- Second Model
  - Clearances (further investigating will be done)
  - Shots on target %
  - Crosses (had negative impact in LDA/log. reg.)
    - We do not know sign of the impact for any of these

Conclusions & Future Work

- Shortcomings:
  - There are of course shortcomings to this analysis. We are limited with the level of data we have, and not every tactical principle can really be measured. Second, every opponent is different, and these results need to be taken with a grain of salt. Lastly, the MLS is a difficult league to model due to the high level of team parity.
- Continuing On:
  - We are continuing to brainstorm different variables to be included in our primary model. We also plan to test smaller models and compare accuracy as a way of comparing metrics at a more secular level. This is another way to compare metrics, other than what we have already done with the coefficients. Lastly, we will do some within class clustering.