Atlanta United Project 2

Jake Federman

2025-03-24

Objective: Find 3 attacking players who stand out as potential targets for MLS clubs

#Reading in the Data

```
events <- readRDS("atlutd_datascientist_project2_eventdata.rds")
schedule2 <- read_csv(file = "atlutd_datascientist_project2_schedule.csv")
```

```
## New names:
## Rows: 111 Columns: 7
## -- Column specification
## -------------------------------------------------------- Delimiter: "," dbl
## (7): ...1, match_id, home_score, away_score, match_week, home_team_id, a...
## i Use `spec()` to retrieve the full column specification for this data. i
## Specify the column types or set `show_col_types = FALSE` to quiet this message.
## * `` -> `...1`
```

```
caps <- read_csv(file = "atlutd_datascientist_project2_player_mins_appearances.csv")
```

```
## New names:
## Rows: 664 Columns: 5
## -- Column specification
## -------------------------------------------------------- Delimiter: "," dbl
## (5): ...1, player_id, team_id, player_season_appearances, player_season_...
## i Use `spec()` to retrieve the full column specification for this data. i
## Specify the column types or set `show_col_types = FALSE` to quiet this message.
## * `` -> `...1`
```

Filter the data down to only attacking player events

```
events <-
  events %>%
  filter(position.name %in% c("Center Forward",
                              "Center Attacking Midfield",
                              "Right Attacking Midfield",
                              "Left Attacking Midfield",
                              "Right Wing",
                              "Left Wing",
                              "Right Center Forward",
                              "Left Center Forward"))
```

Creating new stats

Passes Completed Over Expected

We will create an Outs Above Average style metric for pass completion.

Let X_{ij} be the probability of completing pass i by player j. Let Y_{ij} be the outcome of pass i by player j: 0 if incomplete, 1 if complete Let n be the number of passes player j attempted this season.

Player j's $PCOE$ (passes completed over expected) is equal to the sum of his Y_i - X_i values

We will not include set pieces

```
passes <-
  events %>%
  filter(type.name == "Pass" &
           (is.na(pass.type.name) |
            (!pass.type.name %in%
               c("Kick Off", "Corner", "Free Kick", "Throw-in")))) %>%
  mutate(
    pass_complete = case_when(
    is.na(pass.outcome.name) ~ 1,
    TRUE ~ 0),
    pcoe = pass_complete - pass.pass_success_probability
    ) %>%
  select(player.id, pass_complete, pass.pass_success_probability, pcoe)

pcoe_leaderboard <-
  passes %>%
  group_by(player.id) %>%
  summarize(num_passes = n(),
            pcoe_total = sum(pcoe, na.rm = TRUE),
            pcoe_per_pass = pcoe_total/num_passes) %>%
  arrange(desc(pcoe_total))

#Standardize
standardize <- function(x) {
  mu <- mean(x, na.rm = TRUE)
  sigma <- sd(x, na.rm = TRUE)
  return( (x - mu)/sigma )
}

z_pcoe_leaderboard <-
  pcoe_leaderboard %>%
  mutate(z_pcoe_total = standardize(pcoe_total),
         z_pcoe_per_pass = standardize(pcoe_per_pass)) %>%
  select(1, 2, 5, 6)

z_pcoe_per_pass_leaders <-
  z_pcoe_leaderboard %>%
  arrange(desc(z_pcoe_per_pass))

pcoe_leaderboard %>%
  select(player.id, num_passes, pcoe_total) %>%
  arrange(desc(pcoe_total)) %>%
```

```
  head(10) %>%
  kable(format = "latex", booktabs = TRUE, digits = 3,
        col.names = c("Player", "Passes", "Total PCOE"),
        caption = "Passes Completed Over Expected Leaders",
        align = c("l", "c", "c")) %>%
  kable_styling(position = "center", latex_options = "HOLD_position")
```

Table 1: Passes Completed Over Expected Leaders

Player	Passes	Total PCOE
3109	343	13.083
3814	403	12.650
3629	273	11.031
40784	310	8.180
443805	100	6.138
22695	122	5.815
124863	142	5.589
5261	255	5.477
43133	110	5.251
25289	114	4.952

```
pcoe_leaderboard %>%
  select(player.id, num_passes, pcoe_per_pass) %>%
  arrange(desc(pcoe_per_pass)) %>%
  head(10) %>%
  kable(format = "latex", booktabs = TRUE, digits = 3,
        col.names = c("Player", "Passes", "PCOE per Pass"),
        caption = "Passes Completed Over Expected (per Pass) Leaders",
        align = c("l", "c", "c")) %>%
  kable_styling(position = "center", latex_options = "HOLD_position")
```

Table 2: Passes Completed Over Expected (per Pass) Leaders

Player	Passes	PCOE per Pass
120039	1	0.774
51092	1	0.445
44263	2	0.296
448579	3	0.262
43288	10	0.213
43151	5	0.212
116926	2	0.189
35512	1	0.185
5610	7	0.182
263285	1	0.181

Aggregations Involving Expected Goals

```
shots <-
  events %>%
  filter(type.name == "Shot" & shot.type.name == "Open Play") %>%
  select(player.id, shot.outcome.name, shot.statsbomb_xg, shot.shot_execution_xg,
         shot.shot_execution_xg_uplift)
```

```
#Pre-shot xG leaders
pre_shot_xg_leaders <-
  shots %>%
  group_by(player.id) %>%
  summarize(num_shots = n(),
            total_pre_shot_xg = sum(shot.statsbomb_xg, na.rm = TRUE),
            pre_shot_xg_per_shot = total_pre_shot_xg/num_shots) %>%
  arrange(desc(total_pre_shot_xg))

#Post shot xG leaders
post_shot_xg_leaders <-
  shots %>%
  group_by(player.id) %>%
  summarize(num_shots = n(),
            total_post_shot_xg = sum(shot.shot_execution_xg, na.rm = TRUE),
            post_shot_xg_per_shot =
              total_post_shot_xg/num_shots) %>%
  arrange(desc(total_post_shot_xg))

#xG above expected (post shot xG - pre shot xG)
xgoe_leaders <-
  shots %>%
  group_by(player.id) %>%
  summarize(num_shots = n(),
            total_xgoe = sum(shot.shot_execution_xg_uplift, na.rm = TRUE),
            xgoe_per_shot = total_xgoe/num_shots) %>%
  arrange(desc(total_xgoe))
```

50/50s

```
fifty_fifties <-
  events %>%
  filter(type.name == '50/50') %>%
  select(player.id, `50_50.outcome.id`) %>%
  mutate(fifty_win = case_when(
    `50_50.outcome.id` <= 2 ~ 0,
    TRUE ~ 1
  )) %>%
  group_by(player.id) %>%
  summarize(`50_50s` = n(),
            wins_50_50 = sum(fifty_win, na.rm = TRUE),
            expected_wins_50_50 = `50_50s`/2,
```

```
            wins_above_expected_50_50 = wins_50_50 - expected_wins_50_50) %>%
  ungroup()
```

Dribbles

```
dribbles <-
  events %>%
  filter(type.name == "Dribble") %>%
  mutate(dribble_success = case_when(
    dribble.outcome.id == 8 ~ 1,
    TRUE ~ 0
  )) %>%
  select(player.id, dribble_success) %>%
  mutate(overall_success_rate = sum(dribble_success)/n()) %>%
  group_by(player.id, overall_success_rate) %>%
  summarize(num_dribble_attempts = n(),
            num_dribble_successes = sum(dribble_success, na.rm = TRUE),
            dribble_success_rate = num_dribble_successes/num_dribble_attempts) %>%
  ungroup() %>%
  mutate(expected_dribble_successes = overall_success_rate * num_dribble_attempts,
         dribbles_over_expected =
           num_dribble_successes - expected_dribble_successes) %>%
  select(-overall_success_rate)
```

```
## `summarise()` has grouped output by 'player.id'. You can override using the
## `.groups` argument.
```

Join the new stats together

```
new_stats_leaderboard <-
  pcoe_leaderboard %>%
  inner_join(pre_shot_xg_leaders, by = 'player.id') %>%
  inner_join(post_shot_xg_leaders, by = 'player.id') %>%
  inner_join(xgoe_leaders, by = 'player.id') %>%
  inner_join(dribbles, by = 'player.id') %>%
  left_join(fifty_fifties, by = 'player.id') %>%
  mutate(across(everything(), ~replace_na(., 0)))

new_stats_totals <-
  new_stats_leaderboard %>%
  select(player.id, num_passes, num_shots, num_dribble_attempts,
         `50_50s`, pcoe_total, total_pre_shot_xg,
         total_xgoe, dribbles_over_expected, wins_above_expected_50_50)
```

Combine new leaderboard with minutes played data

```
caps_simple <-
  caps %>%
  mutate(minutes_per_game = player_season_minutes/player_season_appearances) %>%
  select(player_id, player_season_minutes, minutes_per_game) %>%
  rename(player.id = player_id)

combined_stats <-
  new_stats_totals %>%
  inner_join(caps_simple, by = 'player.id')
```

Put key stats on a per 90 minute basis

```
combined_stats_per_90 <-
  combined_stats %>%
  mutate(pcoe_per_90 = 90*(pcoe_total/player_season_minutes),
         xg_per_90 = 90*(total_pre_shot_xg/player_season_minutes),
         xgoe_per_90 = 90*(total_xgoe/player_season_minutes),
         doe_per_90 = 90*(dribbles_over_expected/player_season_minutes),
         wae_50_50 = 90*(wins_above_expected_50_50/player_season_minutes)) %>%
  select(-c(6:10))
```

```
#Take z-scores
z_combined_stats_per_90 <-
  combined_stats_per_90 %>%
  mutate(z_pcoe_per_90 = standardize(pcoe_per_90),
         z_xg_per_90 = standardize(xg_per_90),
         z_xgoe_per_90 = standardize(xgoe_per_90),
         z_doe_per_90 = standardize(doe_per_90),
         z_wae_50_50 = standardize(wae_50_50)) %>%
  select(1:7, 13:17)

z_sum_combined_stats_per_90 <-
  z_combined_stats_per_90 %>%
  group_by(player.id) %>%
  mutate(total_z_score = sum(across(c(starts_with("z_"))))) %>%
  arrange(desc(total_z_score))
```

The best players by total_z_score

```
z_sum_combined_stats_per_90 %>%
  select(player.id, total_z_score) %>%
  head(10) %>%
  kable(format = "latex", booktabs = TRUE, digits = 3,
        col.names = c("Player", "Total Z-Score"),
        caption = "The best players by total z score",
        align = c("l", "c", "c")) %>%
  kable_styling(position = "center", latex_options = "HOLD_position")
```

Table 3: The best players by total z score

Player	Total Z-Score
130625	9.698
4269	5.409
3990	5.257
5207	4.834
386448	4.639
43133	3.997
37649	3.921
127952	3.363
51395	3.336
30421	3.213

Hand Picking the Ideal Targets

I will hand pick the ideal targets taking into account not only total z-score, but also sample size. The top player according to total z score has only played 75 minutes all season across 2 matches, and he has 2 shots. We should not take a chance on him with this little information.

Player 1: Player 4269

```
row_4269 <-
  z_sum_combined_stats_per_90 %>%
  filter(player.id == 4269)

radar_values <-
  row_4269 %>%
  ungroup() %>%
  select(z_pcoe_per_90, z_xg_per_90, z_xgoe_per_90, z_doe_per_90, z_wae_50_50) %>%
  rename(`Passing (-0.09)` = z_pcoe_per_90,
         `Scoring chances (2.31)                ` = z_xg_per_90,
         `Finishing (1.91)` = z_xgoe_per_90,
         `Dribbling (1.08)` = z_doe_per_90,
         `50/50s (0.2)` = z_wae_50_50)

rescaled <-
  as.data.frame(lapply(radar_values, function(x) (x + 3) / 6))

radar_df <-
  rbind(rep(1, 5),
                 rep(0, 5),
                 rescaled)

colnames(radar_df) <-
  colnames(radar_values)
rownames(radar_df) <-
  c("max", "min", "Player")
```

```r
radarchart(radar_df,
           pcol = "blue",
           pfcol = alpha("blue", 0.4),
           plwd = 2,
           cglcol = "grey",
           cglty = 1,
           axislabcol = "grey20",
           vlcex = 0.8,
           title = "Player 4269 Radar Plot")

title(sub = "Total z-score = 5.41 (2nd)", cex.sub = 1)
```

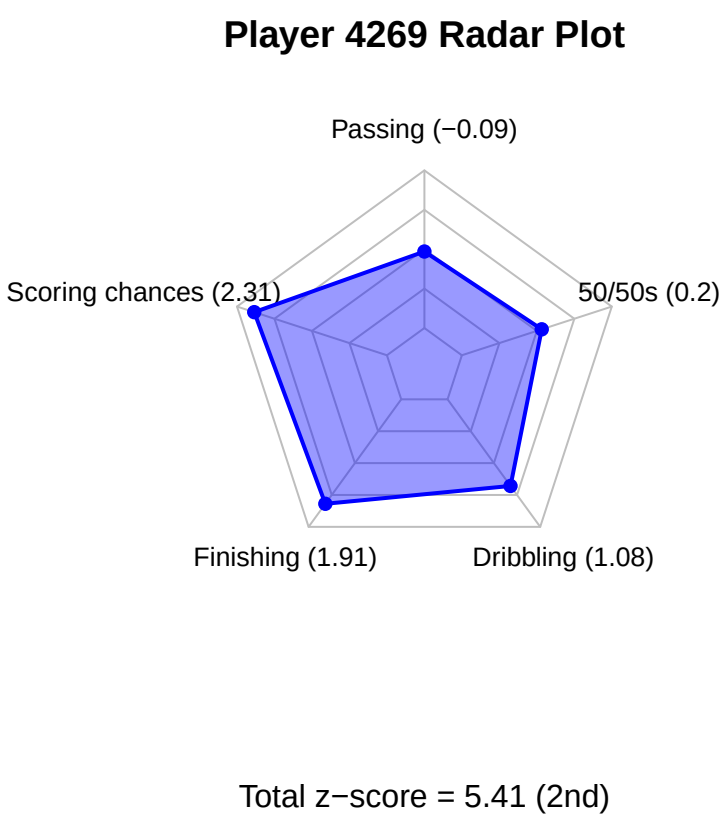

Figure 1: Radar plot for Target 1, player 4269

Scouting for the second and third players:

In searching for suitable transfer targets, we must look for both good quality and great value. As a result, for the next two targets, I will not merely choose the best players, but I will look for ones that are being used as substitutes. If we were to inquire about a player who is playing 95 minutes every match, the team will likely demand a high price, as that player is a crucial part of their organization. Perhaps they will be more willing to sell a player who is on the pitch less, maybe coming off the bench as a sub, allowing us to get a better price.

```r
z_sum_combined_stats_per_90 %>%
  filter(minutes_per_game <= 50 & player_season_minutes >= 250
         & num_passes >= 50 & num_shots >= 5 & total_z_score >= 1.5)
```

```
## # A tibble: 3 x 13
## # Groups:   player.id [3]
##   player.id num_passes num_shots num_dribble_attempts `50_50s`
##       <dbl>      <int>     <int>                <int>    <int>
## 1     52065         57         8                    4        1
## 2     42137         96        10                    6        2
## 3    443805        100         7                    5        0
## # i 8 more variables: player_season_minutes <dbl>, minutes_per_game <dbl>,
## #   z_pcoe_per_90 <dbl>, z_xg_per_90 <dbl>, z_xgoe_per_90 <dbl>,
## #   z_doe_per_90 <dbl>, z_wae_50_50 <dbl>, total_z_score <dbl>
```

Player 2: Player 52065

```r
row_52065 <-
  z_sum_combined_stats_per_90 %>%
  filter(player.id == 52065)

radar_values <-
  row_52065 %>%
  ungroup() %>%
  select(z_pcoe_per_90, z_xg_per_90, z_xgoe_per_90, z_doe_per_90, z_wae_50_50) %>%
  rename(`Passing (1.29)` = z_pcoe_per_90,
         `Scoring chances (1.04)                ` = z_xg_per_90,
         `Finishing (0.329)` = z_xgoe_per_90,
         `Dribbling (0.582)` = z_doe_per_90,
         `50/50s (-0.784)` = z_wae_50_50)

rescaled <-
  as.data.frame(lapply(radar_values, function(x) (x + 3) / 6))

radar_df <-
  rbind(rep(1, 5),
                 rep(0, 5),
                 rescaled)

colnames(radar_df) <-
  colnames(radar_values)
rownames(radar_df) <-
  c("max", "min", "Player")

radarchart(radar_df,
           pcol = "blue",
           pfcol = alpha("blue", 0.4),
           plwd = 2,
           cglcol = "grey",
           cglty = 1,
           axislabcol = "grey20",
```

```
           vlcex = 0.8,
           title = "Player 52065 Radar Plot")

title(sub = "Total z-score = 2.45 (1st among players with more than 200 min
      on the season and less than 45 minutes per game)", cex.sub = 1)
```

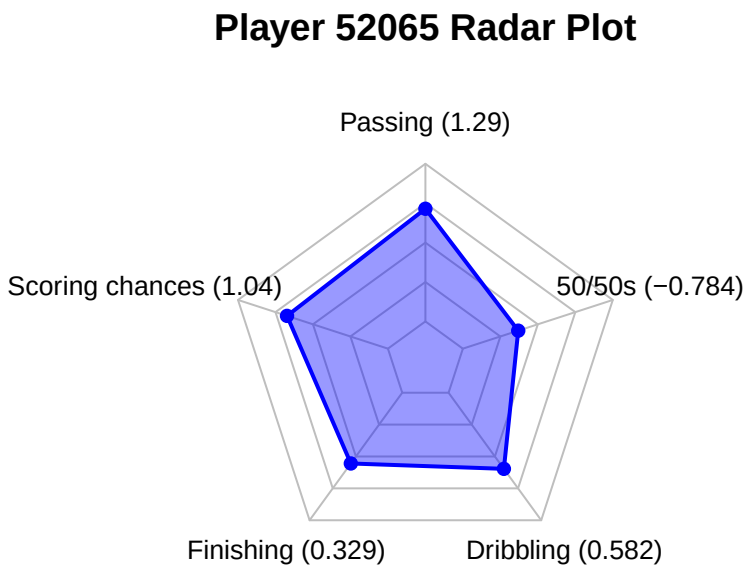

Figure 2: Radar plot for Target 2, player 52065

Player 3: Player 42137

```
row_42137 <-
  z_sum_combined_stats_per_90 %>%
  filter(player.id == 42137)

radar_values <-
  row_42137 %>%
  ungroup() %>%
  select(z_pcoe_per_90, z_xg_per_90, z_xgoe_per_90, z_doe_per_90, z_wae_50_50) %>%
  rename(`Passing (0.390)` = z_pcoe_per_90,
         `Scoring chances (-0.212)                ` = z_xg_per_90,
         `Finishing (1.71)` = z_xgoe_per_90,
         `Dribbling (0.019)` = z_doe_per_90,
```

```
        `50/50s (0.196)` = z_wae_50_50)

rescaled <-
  as.data.frame(lapply(radar_values, function(x) (x + 3) / 6))

radar_df <-
  rbind(rep(1, 5),
                  rep(0, 5),
                  rescaled)

colnames(radar_df) <-
  colnames(radar_values)
rownames(radar_df) <-
  c("max", "min", "Player")

radarchart(radar_df,
           pcol = "blue",
           pfcol = alpha("blue", 0.4),
           plwd = 2,
           cglcol = "grey",
           cglty = 1,
           axislabcol = "grey20",
           vlcex = 0.8,
           title = "Player 42137 Radar Plot")

title(sub = "Total z-score = 2.10 (2nd among players with more than 200 min
      on the season and less than 45 minutes per game)", cex.sub = 1)
```

Extra PCOE leaderboard

```
combined_stats_per_90 %>%
  select(player.id, pcoe_per_90) %>%
  arrange(desc(pcoe_per_90)) %>%
  head(10) %>%
  kable(format = "latex", booktabs = TRUE, digits = 3,
        col.names = c("Player", "PCOE per 90 Minutes"),
        caption = "PCOE per 90 Minutes Leaderboard",
        align = c("l", "c")) %>%
  kable_styling(position = "center", latex_options = "HOLD_position")
```

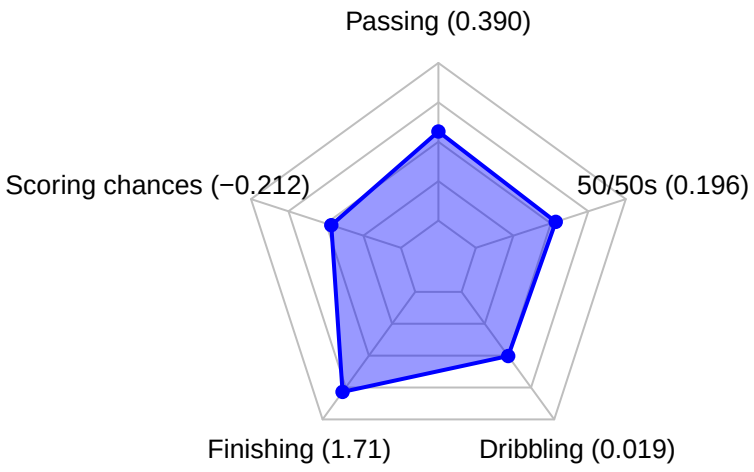

Figure 3: Radar plot for Target 3, player 42137

Table 4: PCOE per 90 Minutes Leaderboard

Player	PCOE per 90 Minutes
274630	4.028
443805	2.116
3109	1.445
51543	1.267
25289	1.250
3629	1.236
53767	1.202
3814	1.164
130625	1.137
10755	1.083