PREDICTING PLAYER PASSES

Statistical Modeling & Market Efficiency Analysis in the Premier League

Kian Peters Mondry | Austin Huang | Christopher Morales | Suyesh Niraula

TABLE OF CONTENTS

Introduction and Scope

Model Goal

Narrowing focus to **Bootstrapping & Confidence Intervals** for predicting player passes attempted in a single match.

Team Select

Focused on **Liverpool FC** due to stable possession stats and less random tactical outcomes.

Benchmark

Comparing internal results against **Market Efficiency** (Betting Lines) and real match outcomes.

Liverpool F.C. - Background

Possession Dominance

- Liverpool FC holds the highest average possession in the Premier League
- Results in high volume touches and passes across the entire team.

The Betting Line: Bookmakers offer "Over/Under" lines on player passes

- Goal is to outperform these professional models using logic and advanced statistics

STEP 1: Finding the Factors

After conducting research and reading other studies that analyzed certain box statistics in the field of soccer we decided on factors based on:

Historical Form

Previous passes attempted and recent 5-game rolling averages.

Venue & Rival

Home vs. Away dynamics and opponent's average possession rate.

Live Match State

Red cards, injuries, expected minutes, and Half-Time (HT) results.

Examples of How Factors influence player passes

- **Opponent Red Cards:** Increases team possession, leading to higher individual pass counts.
- **Losing at Half Time:** Often Forces a higher "keep ball" urgency, increasing possession in the second half - especially for midfielders bringing the ball up the field.
- **Positioning:** Defenders (like Van Dijk) and CDMs see significantly more build-up ball than attackers.
- **Minutes Played:** Expected playtime (90' vs 70') is the primary multiplier for volume prediction.

Collecting Data

2025/26 Season Focus

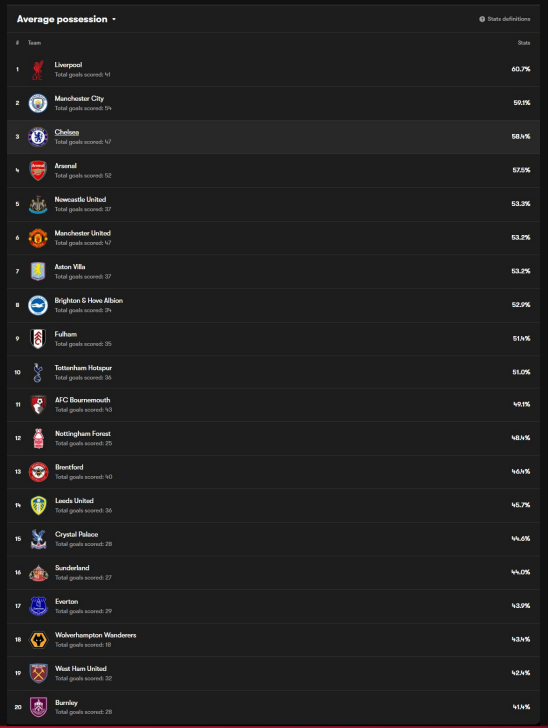
Utilized Fotmob and Football-Data.co.uk to extract all factors from the past 26 games.

Stability over Volume: Excluded previous seasons to avoid "tactical noise" caused by different players, roles, and managerial instructions.

Will overcome less data with bootstrapping in our models which sample with replacement

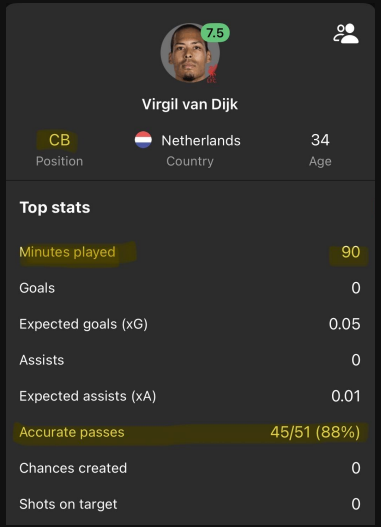
Examples of what each data looks like before extracting and grouping together

Avg possession factor for teams this season:



#	Team	Stats
1	Liverpool — Total goals scored: 41	60.7%
2	Manchester City — Total goals scored: 54	59.1%
3	Chelsea — Total goals scored: 47	58.1%
4	Arsenal — Total goals scored: 52	57.5%
5	Newcastle United — Total goals scored: 37	53.3%
6	Manchester United — Total goals scored: 47	53.2%
7	Aston Villa — Total goals scored: 37	53.2%
8	Brighton & Hove Albion — Total goals scored: 34	52.9%
9	Fulham — Total goals scored: 35	51.1%
10	Tottenham Hotspur — Total goals scored: 36	51.0%
11	AFC Bournemouth — Total goals scored: 43	49.1%
12	Nottingham Forest — Total goals scored: 25	48.1%
13	Brentford — Total goals scored: 40	46.1%
14	Leeds United — Total goals scored: 36	45.7%
15	Crystal Palace — Total goals scored: 28	44.6%
16	Sunderland — Total goals scored: 27	44.0%
17	Everton — Total goals scored: 29	43.9%
18	Wolverhampton Wanderers — Total goals scored: 18	43.1%
19	West Ham United — Total goals scored: 32	42.1%
20	Burnley — Total goals scored: 28	41.1%

Position, minutes and passes attempted for 1 game factors:



Example of how we list all the data together before we make data frame:

Game 24 - Home vs Newcastle (53.3% avg possession team). up 2-1 at halftime. Played full game. No red cards and did not get injured. 51 passes attempted. Played cb. again had midweek game but did not affect PT

Game 23 - away vs Bournmouth (49.1% possession team). Down 2-1 at half time (imp as maybe leads to them attacking more) no red cards on teams and player wasn't injured. Played cb. 73 passes attempted.

Game 22 - Home vs Burnly (41.4%) possession team. Up 1-0 at half time. No red cards on teams and no injury. Played full game. Played cb. 98 passes attempted.

Game 21 - Away vs Arsenal (57.5% possession team). 0-0 at half. No reds and no injury. full game played. Played cb. 83 passes attempted

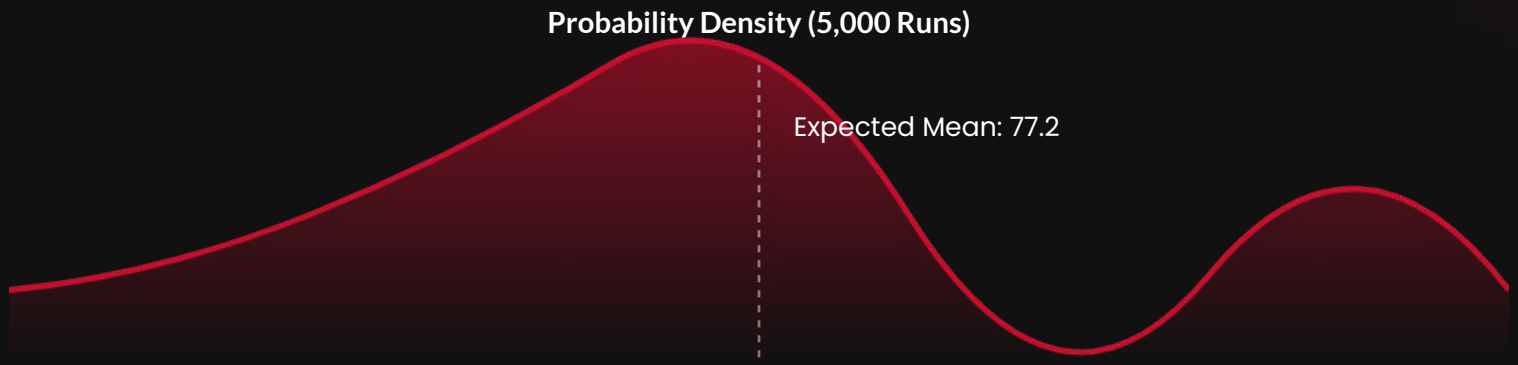
Game 20 - away vs fulham (51.4% possession team). Down 1-0 at half. No reds and no injury. Full game played. played cb. 87 passes attempted

Game 19 - home vs Leeds (45.7% possession team). 0-0 at half. no reds and no injury. Playedfull game. Played cb. 91 passesattempted.

Game 18 - home vs wolves (43.4% possession team). 2-0 up at the half. No reds and no injury. Full game played. Played cb. 102 passes attempted.

Game 17 - away vs Totenham (51% possession team. 0-0 at half. 2 reds for opponent. 1 early on in 33 min and one at the 90+3 (the second barely affects data but first one led to lopsided possession for liverpool. No injury. Played full game. Played cb. 79 passes

MODELING: BOOTSTRAPPING



Bootstrapping over 5,000 runs from trained data compensates for the small sample size (n=26) and generate robust 95% confidence intervals.

- It bootstraps a data set that contains all the factors in a linear model from all games of the season which trains the prediction and makes a new one based on the new factors for the upcoming game.

MODELING: Methods

Bootstrapping will then be used to:

Predict for overall regular game and for different scenarios:

- Ex: predict passes attempted for when winning, losing or drawing at half time

Compare to real results:

- Compare our model prediction to actual result and our prediction to market efficiencies lines

Do multiple tests:

- Will run comparisons on several games as this would ideally be a longitudinal study that compares model accuracy over a long period of time to reduce possibility of random chance which is common in sports
  - (however can only do 3 due to timing of presentation)

Test 1: VIRGIL VAN DIJK

Van Dijk vs. Nottingham Forest

For the first game we are testing the prediction on Van Dijk versus Nottingham forest away (relatively med-low possession based team)

Consistency Factor: Van Dijk played 90 minutes in 100% of matches. His position is fixed at CB, simplifying two our parameters (position and minutes).

Red Cards: actually played a negligible role in Liverpool games and did not affect Van Dijks passes -> removed as a factor

Half Time result: Halftime results had a difference but not much of a change (will see in final model results)

Home/Away: home versus away games did show noticeable differences in his passes attempted.

Form: His passes attempted in past couple games has been lower than the average for season. May indicate shift in team play style so expect lower end of our confidence interval

Live Game Expectation: Liverpool have struggled versus Nottingham Forest historically and lost 3-0 to them earlier this season. Expect competitive game so maybe drawing at half time

Van Dijk Final Data Frame

Note: the order is backwards. The last couple games are first in data frame

Passing Performance Table

Passes	Minutes	Opponent Poss %	Home	HT State (-1=L,0=T,1=W)	Passes per 90	Passes per Min
58.00	90.00	44.00	0.00	0.00	58.00	0.64
45.00	90.00	58.10	1.00	0.00	45.00	0.50
51.00	90.00	53.30	1.00	1.00	51.00	0.57
73.00	90.00	49.10	0.00	−1.00	73.00	0.81
98.00	90.00	41.40	1.00	1.00	98.00	1.09
83.00	90.00	57.50	0.00	0.00	83.00	0.92
87.00	90.00	51.40	0.00	−1.00	87.00	0.97
91.00	90.00	45.70	1.00	0.00	91.00	1.01
102.00	90.00	43.40	1.00	1.00	102.00	1.13
79.00	90.00	51.00	0.00	0.00	79.00	0.88
65.00	90.00	52.90	1.00	1.00	65.00	0.72
64.00	90.00	45.70	0.00	0.00	64.00	0.71
97.00	90.00	44.00	1.00	0.00	97.00	1.08
80.00	90.00	42.70	0.00	0.00	80.00	0.89
106.00	90.00	48.40	1.00	−1.00	106.00	1.18
47.00	90.00	58.10	0.00	−1.00	47.00	0.52
87.00	90.00	53.20	1.00	1.00	87.00	0.97
76.00	90.00	46.40	0.00	−1.00	76.00	0.84
68.00	90.00	53.20	1.00	−1.00	68.00	0.76
60.00	90.00	58.40	0.00	−1.00	60.00	0.67
105.00	90.00	44.60	0.00	−1.00	105.00	1.17
67.00	90.00	43.90	1.00	1.00	67.00	0.74
105.00	90.00	41.40	0.00	0.00	105.00	1.17
81.00	90.00	57.50	1.00	0.00	81.00	0.90
62.00	90.00	53.30	0.00	1.00	62.00	0.69
74.00	90.00	49.10	1.00	1.00	74.00	0.82

Test 1: VAN DIJK PREDICTION

Full model predictions

```
Rate of passing: 0.8582104
Rate of passing CI: 0.7646101 to 0.9543477
Expected number of passes for full game: 77.23894
CI for expected full-game passes: 68.81491 to 85.89129

 Halftime Scenario:
Prob Win HT: 0.3076923
Prob Tie HT: 0.3846154
Prob Lose HT: 0.3076923

Winning at HT:
Expected passes: 74.68894
CI: 66.26491 to 83.34129

Tied at HT:
Expected passes: 77.23894
CI: 68.81491 to 85.89129

Losing at HT:
Expected passes: 76.68894
CI: 68.26491 to 85.34129

Weighted expected passes: 76.28509
Weighted expected passes CI: 67.86106 to 84.93745
```

77

PREDICTED MEAN

68-86

95% CONF. INTERVAL

76

WEIGHTED EXPECTED

Expectations: regardless of half time result, model is centered around expecting ~76 passes but due to recent form we predict, he will likely get around the **Lower End** of the interval (~70-74).

Betting models online: Models had Van Dijk's line for over/under at 77.5 passes attempted (very similar to ours). Still given our info our end conclusion of around low 70's would lead us to predicting the under

77.5

Over/Under for betting Models

Actual Results & vs MARKET EFFICIENCY

74

Actual Result

Low 70’s

Predicted Result from model

77.5

Correct decision from betting line

Conclusion: Our assessment of low 70’s was not far off from the actual result of 74 and well within the confidence interval

- Our logic-based weighting successfully identified the "Under" value, beating market efficiency by identifying form discrepancies.
- shows a slight edge on our model and logic beating market efficiency (at least for this game)

TEST 2: RYAN GRAVENBERCH

Gravenberch vs West Ham

Next prediction is for a Home game against West ham (who have second lowest average possession in the league

Position Volatility: Gravenberch (CDM/CM) shows much higher data variance than defenders. His role is susceptible to early substitutions in blowouts and game management.

Factor Shift: Winning at HT resulted in significantly lower passes compared to Van Dijk, where match state barely affected volume.

Red cards, opponent possession and Home/away factors: Pretty much the same as Van Dijks

- Red cards again removed as factor and others are regarded as important factors

Minutes: Had games where didn't play full game or game at all so we have to be careful converting minutes and passes (missing games were removed)

- Expect a full 90 for this game though as no CL game midweek

PT2

Role Evolution

Recent 4 games: 40.25 avg passes. Rest of season: 60.3 avg. Indicates a massive role change in the pivot.

Weighting Solution

Applied **5:1 Sampling Weight** to recent form during bootstrapping to bias the model toward current tactical reality.

Live game Expectation

Playing a team that they almost always beat and who is currently in relegation. Expect liverpool to win this game -> lean towards winning at half model

GRAVENBERCH Final Data Frame

Note: The last couple games are first in data frame

Passing Performance Table

Passes	Minutes	Opponent Poss %	Home	HT State (-1=L,0=T,1=W)	Passes per 90	Passes per Min
31.00	90.00	48.40	0.00	0.00	31.00	0.34
54.00	90.00	44.00	0.00	0.00	54.00	0.60
39.00	90.00	58.10	1.00	0.00	39.00	0.43
37.00	90.00	53.30	1.00	1.00	37.00	0.41
68.00	90.00	49.10	0.00	−1.00	68.00	0.76
71.00	78.00	41.40	1.00	1.00	81.92	0.91
55.00	90.00	57.50	0.00	0.00	55.00	0.61
75.00	90.00	51.40	0.00	−1.00	75.00	0.83
65.00	90.00	45.70	1.00	0.00	65.00	0.72
67.00	90.00	43.40	1.00	1.00	67.00	0.74
69.00	90.00	51.00	0.00	0.00	69.00	0.77
45.00	90.00	52.90	1.00	1.00	45.00	0.50
49.00	90.00	45.70	0.00	1.00	49.00	0.54
68.00	90.00	44.00	1.00	0.00	68.00	0.76
65.00	90.00	42.70	0.00	0.00	65.00	0.72
70.00	90.00	48.40	1.00	−1.00	70.00	0.78
58.00	90.00	58.10	0.00	−1.00	58.00	0.64
52.00	90.00	53.20	1.00	1.00	52.00	0.58
47.00	62.00	53.20	1.00	−1.00	68.23	0.76
58.00	90.00	58.40	0.00	−1.00	58.00	0.64
77.00	90.00	44.60	0.00	−1.00	77.00	0.86
56.00	90.00	43.90	1.00	1.00	56.00	0.62
94.00	90.00	41.40	0.00	0.00	94.00	1.04
57.00	90.00	57.50	1.00	0.00	57.00	0.63
57.00	90.00	53.30	0.00	1.00	57.00	0.63

Test 2: Gravenberch PREDICTION

Full model predictions

```
Rate of passing (per min): 0.6644956
Rate of passing CI: 0.4912559 to 0.8337967
expected passes (Full Game): 59.80461
CI for expected full-game passes: 44.21303 to 75.0417

 Halftime Scenario
Prob Win HT: 0.2926829
Prob Tie HT: 0.5365854
Prob Lose HT: 0.1707317

Winning at HT:
Expected passes: 59.53304
CI: 43.94147 to 74.77014

Tied at HT:
Expected passes: 59.80461
CI: 44.21303 to 75.0417

Losing at HT:
Expected passes: 77.86933
CI: 62.27776 to 93.10643

Weighted expected passes: 62.80935
Weighted expected passes CI: 47.21777 to 78.04644
```

59

PREDICTED MEAN

44–75

95% CONF. INTERVAL

62

WEIGHTED EXPECTED
(heavily influenced by losing at HT)

Expectations: Liverpool are very likely to win this game (odds have them at 71% win and 14% draw -> assume we will very likely not get losing at HT predictions. Also with form of last 4 games being around 40, a lower assumption is more comfortable. Expect around mid 50's.

Betting models online: Models had Gravenberch line for over/under at 63.5 passes attempted then later at 61.5 (very similar to ours). Still despite our weighted expected passes at 62, given our info we suggest the under.

61.5

Over/Under for betting Models

Actual Results & vs MARKET EFFICIENCY

40

Actual Result

Low 50’s

Predicted Result from model

61.5

Correct decision from betting line

Conclusion: the actual result of 40 was way lower than our prediction of low 50’s, and was also outside most of our confidence intervals

Note: he played 86 min not full game, and this game was a blowout: 3-0 for liverpool by HT (which likely lowers his passes a lot due to less of a need to keep ball)

- May suggest outlier or we failed to properly account for Gravenberches midfield volatility, especially in new role - Note that he still performed close to his average of past 4 games (40.25) but he played a much lower possession team that usual so it should’ve been higher

Regardless, our logic and stat-based weighting still successfully identified the "Under" value, beating market efficiency by identifying form discrepancies better than they did

- Still shows a slight edge on our model and logic beating market efficiency

TEST 3: VAN DIJK vs Wolves F.C.

Van Dijk vs Wolves

Next prediction is for a away game against Wolves (who have third lowest average possession in the league

Specific Match State: Wolves are a "low block" team. Historical data suggests CB's Liverpool keeps the ball significantly more when they are **Winning at HT** vs. this archetype.

Weighting: Increased recent form weight to **10x** to capture recent games even more for form weighting

Live Game Expectation: Note that wolves are bottom of the table, but have in past 3 games tied to the best team in league, and beat third place despite liverpool being heavy favorites by league position (5th versus 20th)

- Liverpool have also been poor in recent games so Wolves could make it a competitive game and liverpool wont have a blowout victory

- Most things are same from last Van dijk data set in test 1 (red cards , home vs away passes, recent form being lower, position always at CB, and full minutes every game) so same cleaning and factors as test 1
    - Main difference, added a weight factor for form, (increased it slightly to 10 instead of 5 from test 2)
- Another difference: With form weighting it shows half time result has more of an effect on his end passes (will see in result)
    - Last time difference half time result barely mattered for VVD but with new weighting of past 5 it shows big difference in scenarios affected passes attempted

Van Dijk vs Wolves Final Data Frame

Note: The last couple games are first in data frame

Passing Performance Table

Passes	Minutes	Opponent Poss %	Home	HT State (-1=L, 0=T, 1=W)	Passes per 90	Passes per Min
64.00	90.00	42.40	1.00	1.00	64.00	0.71
74.00	90.00	48.40	0.00	0.00	74.00	0.82
58.00	90.00	44.00	0.00	0.00	58.00	0.64
45.00	90.00	58.10	1.00	−1.00	45.00	0.50
51.00	90.00	53.30	1.00	0.00	51.00	0.57
73.00	90.00	49.10	0.00	1.00	73.00	0.81
98.00	90.00	41.40	1.00	1.00	98.00	1.09
83.00	90.00	57.50	0.00	0.00	83.00	0.92
87.00	90.00	51.40	0.00	1.00	87.00	0.97
91.00	90.00	45.70	1.00	1.00	91.00	1.01
102.00	90.00	43.40	1.00	1.00	102.00	1.13
79.00	90.00	51.00	0.00	0.00	79.00	0.88
65.00	90.00	52.90	1.00	0.00	65.00	0.72
64.00	90.00	45.70	0.00	0.00	64.00	0.71
97.00	90.00	44.00	1.00	1.00	97.00	1.08
80.00	90.00	42.70	0.00	0.00	80.00	0.89
106.00	90.00	48.40	1.00	1.00	106.00	1.18
47.00	90.00	58.10	0.00	−1.00	47.00	0.52
87.00	90.00	53.20	1.00	−1.00	87.00	0.97
76.00	90.00	46.40	0.00	1.00	76.00	0.84
68.00	90.00	53.20	1.00	−1.00	68.00	0.76
60.00	90.00	58.40	0.00	0.00	60.00	0.67
105.00	90.00	44.60	0.00	1.00	105.00	1.17
67.00	90.00	43.90	1.00	0.00	67.00	0.74
105.00	90.00	41.40	0.00	1.00	105.00	1.17
81.00	90.00	57.50	1.00	1.00	81.00	0.90
62.00	90.00	53.30	0.00	−1.00	62.00	0.69
74.00	90.00	49.10	1.00	1.00	74.00	0.82

Test 3: Wolves Match Prediction

Full model predictions

```
Rate of passing (per min): 0.7731745
Rate of passing CI: 0.6962801 to 0.8639358
expected passes (Full Game): 69.58571
CI for expected full-game passes: 62.66521 to 77.75422

 Halftime Scenario
Prob Win HT: 0.3013699
Prob Tie HT: 0.5068493
Prob Lose HT: 0.1917808

Winning at HT:
Expected passes: 85.53043
CI: 78.60993 to 93.69894

Tied at HT:
Expected passes: 69.58571
CI: 62.66521 to 77.75422

Losing at HT:
Expected passes: 57.66679
CI: 50.74629 to 65.8353

Weighted expected passes: 72.10515
Weighted expected passes CI: 65.18465 to 80.27366
```

69

PREDICTED MEAN

62–77

95% CONF. INTERVAL

72

WEIGHTED EXPECTED
(heavily influenced by winning at HT)

Expectations: Liverpool should have a competitive game so focus prediction more towards tied at half time result. Leaning towards a mid to high 60's which form would also align with. However, his passes could easily be higher if liverpool have comfortable lead

Betting models online: Models had Van Dijk line for over/under at 80.5 passes. Since we expect competitive game not an easy win versus low block, our model and logic suggests the under

80.5

Over/Under for betting Models

Actual Results & vs MARKET EFFICIENCY

63

Actual Result

mid 60's

Predicted Result from model

80.5

Correct decision from betting line

Conclusion: the actual result of 63 was close to our prediction of mid 60's, and was also within most of our confidence intervals

- Market is likely not accounting for recent form as good as we are, and is making high predictions due to rest of season data, or assumed a comfortable liverpool win
- our logic and stat-based weighting successfully identified the "Under" value, beating market efficiency by identifying form discrepancies better than they did
    - Continues to show a slight edge on our model and logic beating market efficiency

DISCUSSION & RISK FACTORS

- **Sample Size of results:** A longitudinal study across multiple seasons would further validate the Bootstrapping robustness.
- **Factors & Role Evolution:** are there any other underlying factors that could give us more precise calculations? Ex) Assessing role changes remains the biggest challenge; weighting recent form is the best current proxy but there may be better solutions.
- **Limitation: Competition Scope:** Champions League data doesn't carry over well due to varying possession archetypes across Europe. We can only due premier league fixtures right now
- **Limitation: Pure Randomness:** Early injuries or red cards create "Black Swan" events that no statistical model can perfectly quantify. -> is another reason we need more tests to find stable conclusions and make sure random events can be treated more as outliers

Conclusion

- We had 2/3 results that were close to what we were predicting. both times the success was for the Centre back Van Dijk while the worse result was the midfielder Ryan Gravenberch
    - The one failure could be due to randomness of blowout or failure to properly account for new role in midfield for team

- 3/3 times we bested the Market Models showing our model + logic can give you an edge in out predicting Market model
    - Probability To get it right 3/3 times in supposed “even lines” created by Market is ⅛ or .125. Not statistically significant so we need more tests before it can be conclusive but good start to show a slight “edge”

- There are 9 games left in the season, even if this class will be well over with by the time they are all played, our group will continue to predict results to increase the sample size to determine more conclusive results.

Q & A