Kabir Alvaro Hinduja-Obregons

Data Science Final Project

'To Boot or not too Boot'

!!!!!!

Motivation - Football

StatsBomb Conference 2024

Boot It: A Pragmatic Alternative To Build Up Play

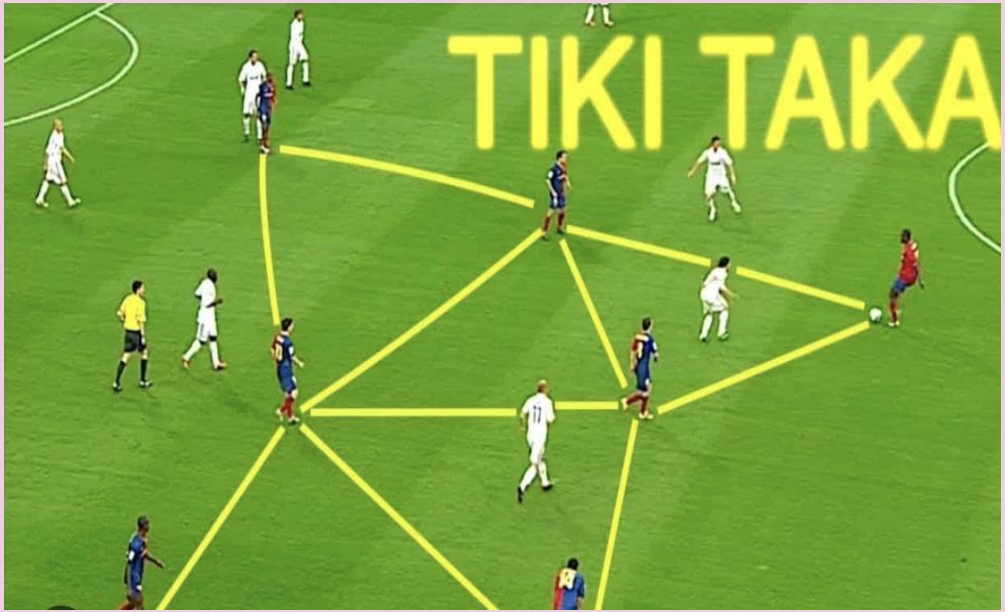
Lorenzo Cascioli,[1] Max Goldsmith,[2] Luca Stradiotti,[1] Maaike Van Roy,[1] Pieter Robberechts,[1] Maxim Wouters[2] and Jesse Davis[1]

[1] *KU Leuven, Dept. of Computer Science; Leuven.AI, B-3000 Leuven, Belgium*

[2] *Royal Belgian Football Association, RBFA Knowledge Centre, B-1020 Brussels, Belgium*

SCHOOL	PTS	PCL RECORD	PCT.	STREAK	HOME	AWAY	NEUTRAL
Father Judge	33	11-0	1.000	L1	7-0	6-0	0-1
La Salle College *	30	10-2	.833	L1	7-1	5-1	2-1
Archbishop Ryan	27	9-2	.818	L1	6-1	4-1	1-2
Archbishop Wood	21	7-4	.636	L1	5-1	2-4	1-0
Conwell-Egan	21	7-4	.636	L1	4-1	3-5	1-0
Lansdale Catholic	19	6-4-1	.591	L1	2-2-1	3-3	1-0
Roman	16	5-6-1	.458	L1	2-2-1	3-3	0-1
St. Joseph's Prep	10	3-7-1	.318	L5	2-4	1-3-1	0-1
Archbishop Carroll	9	3-8	.273	L1	2-3	1-4	0-1
Devon Prep	7	2-8-1	.227	L4	1-4	0-4-1	1-0
Cardinal O'Hara	6	2-9	.182	L6	1-3	0-6	1-0
Bonner & Prendie	0	0-11	.000	L11	0-5	0-5	0-1

Scientific Question: Can a linear regression model better inform players when to 'Boot it' vs trying to play out of the back?



Data and Methods

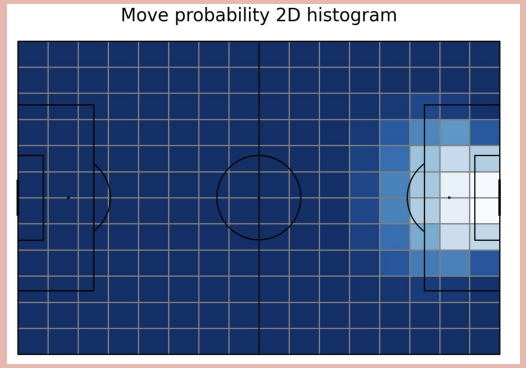
I used free StatsBomb Event and Tracking Data for the WSL between 2018-2021!

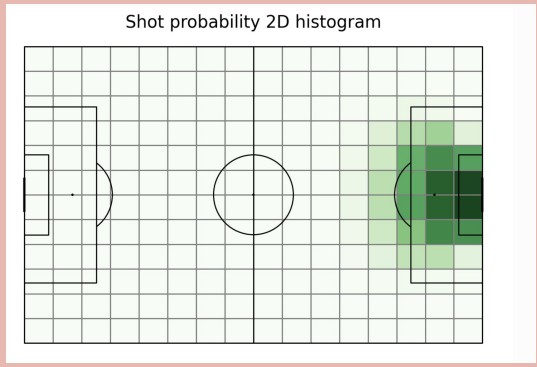
I soon realized there was a reason the data was free, and why the paper I had read was rather straightforward.

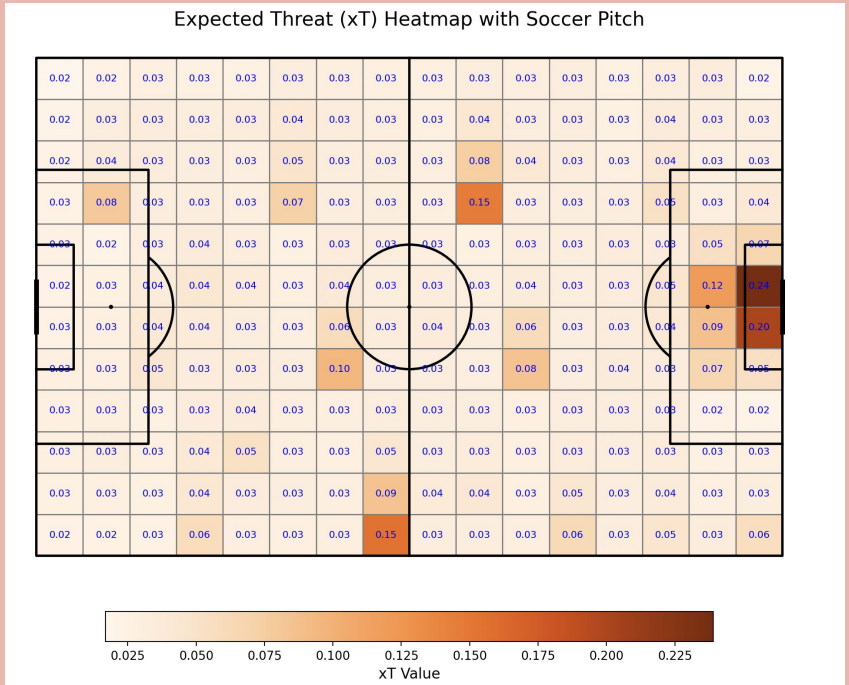
I had ALOT of data I had to generate myself.

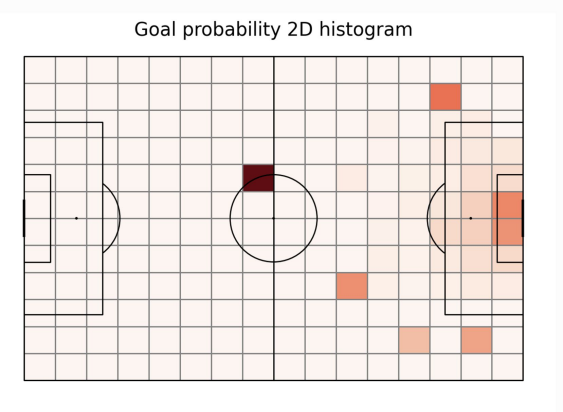
Expected Threat

(some graphs by SoccerMatics on GitHub)

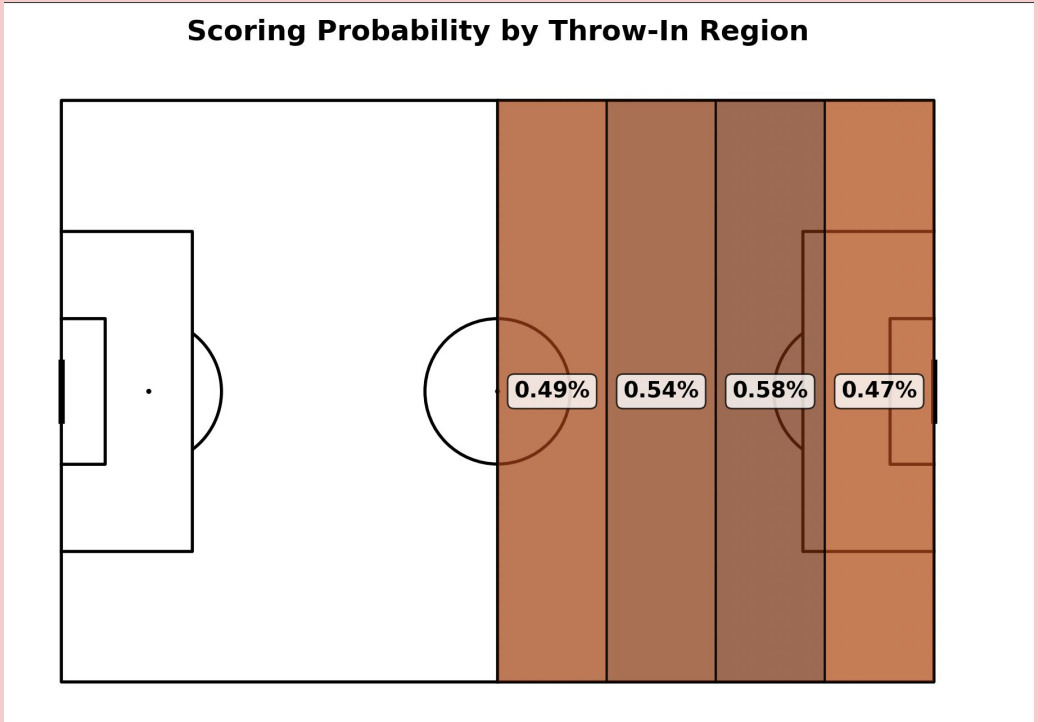




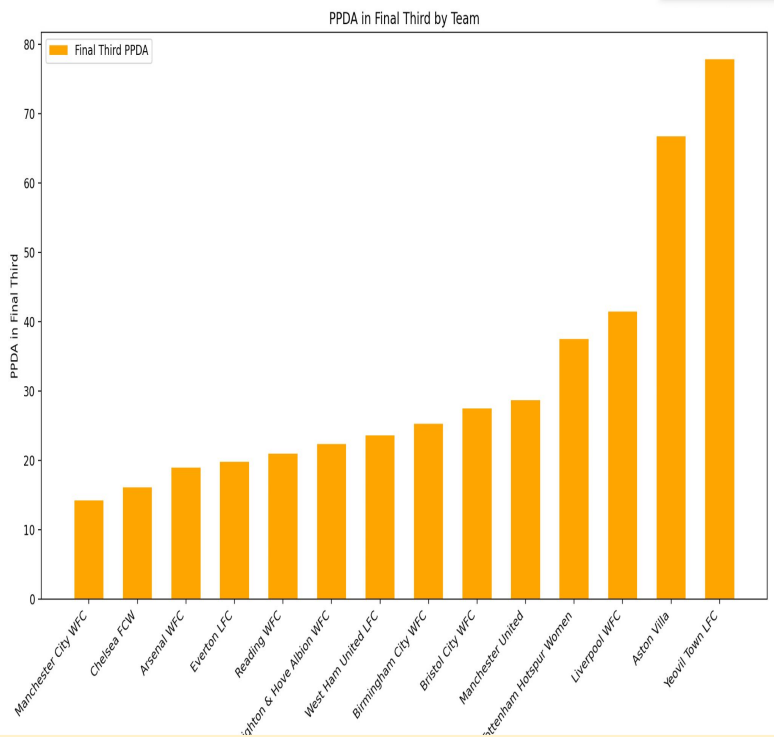




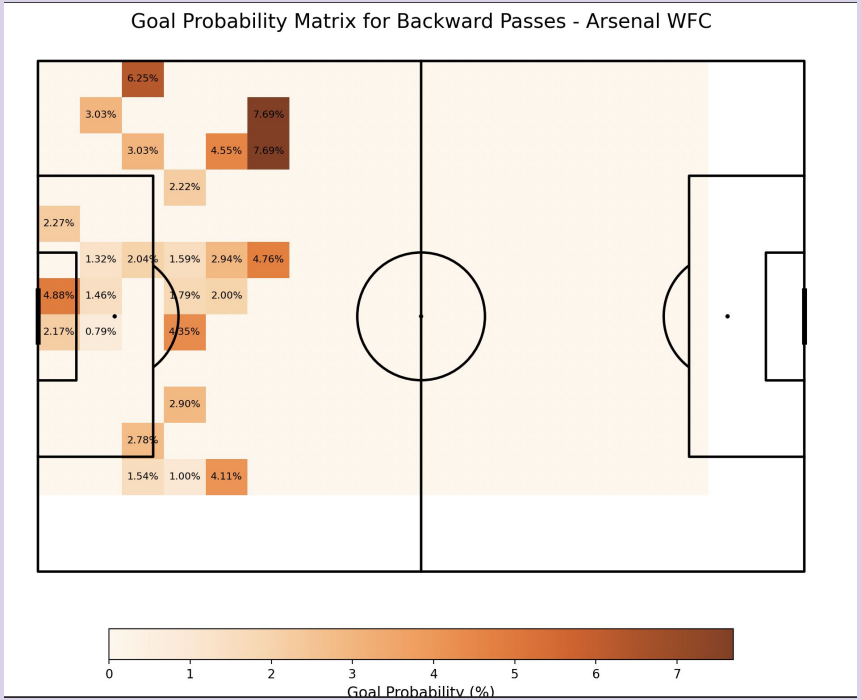
xT when pressing an Opponent's throw in. (What happens after we 'Boot it')



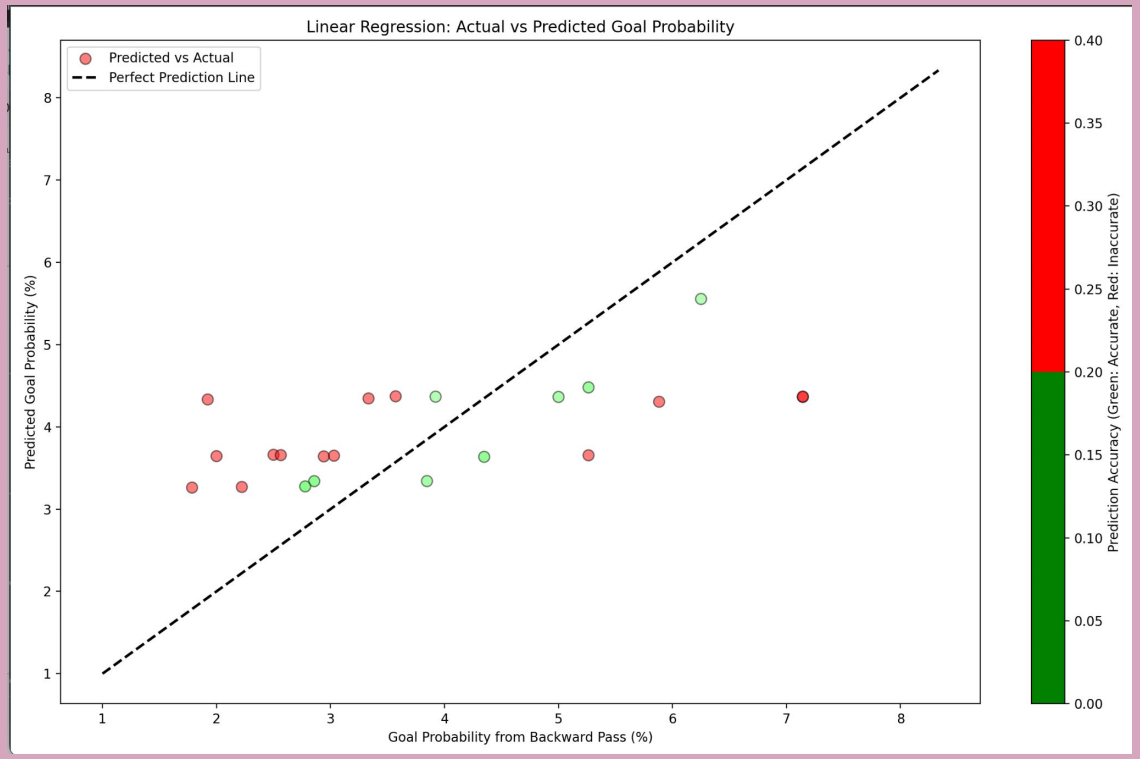
Team Specific Stats: Average Possession, Pressures Per Defensive Action, Average league table finish, and Goals Scored Per Game.



My Label - Each teams likelihood of scoring from a back pass in the build up phase



Results (Finally!) - Linear Regression



Takeaways

Less possession -> more valuable a back pass is.

The more you are pressed -> the more valuable a back pass is.

```
Linear Regression Model Evaluation:
Mean Squared Error: 1.8322
R-squared: 0.3123

Model Coefficients:
xT: -1.3748
Average Possession (%): -0.2096
APF: -0.0483
Average League Position: -0.2129
Average Goals Per Game: 0.9296
```

My Limits!

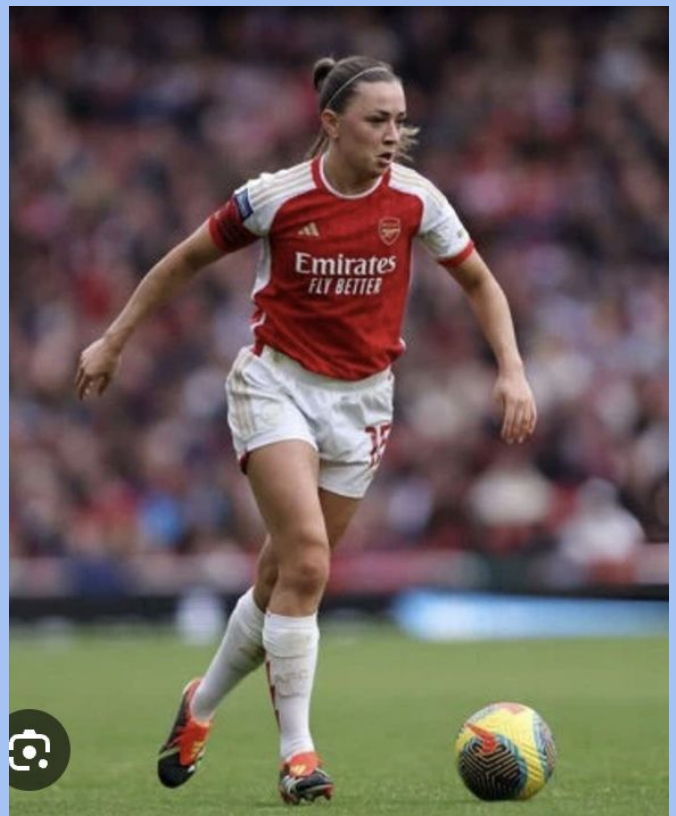
DataSet very small for a football dataset.

Both in terms of what type of data was available (which is why I had to generate so much of it on my own)

And the amount of data there was (3 seasons of 1 league compared to 100's of seasons of data model makers have access too)

Limited historical data for the womens game as well.

Putting my Mid Model to the test! Data Science vs Katie McCabe
Both with a score of 0.79 missed goals over the same 263 back pass situations she was in.



Conclusion and Future Work

My main takeaways are:

- I can do anything I set my mind to!
- …….(But maybe not very well straight away)
- Find more data for sports analytics projects
- Yeovil Town FC is not very good

If I had 6 months I would:

- Create better team specific metrics than Possession and Average Pressure Faced
- Create an Expected Threat Model that includes states when the opponent has the ball

One cool thing…