DS4EPL

Introduction to data science

Group DS4EPL:

- Nguyễn Trần Khang
- Lê Đình Hiếu
- Dương Minh Hoàng
- Lê Anh Quân

Outline

- Introduction
- Method
  - Data
  - Process
  - Approach
- Result
- Summary

Introduction

- The **Premier League (EPL)** is the **most watched football league** in the world
- Analysis EPL
    - Player
    - Match
    - Club

Dataset

- Dataset crawled from: https://www.premierleague.com
- From season 93/94 - now
- Comprise:
    - Player stats: appearances, goals, saves, …
    - Match stats: result, passes, …
    - Club stats: goals, penalty, …
    - Rank

Process

- Crawl data
- Clean and Preprocess data
- Visualize
- Model and evaluate

NaN problems:

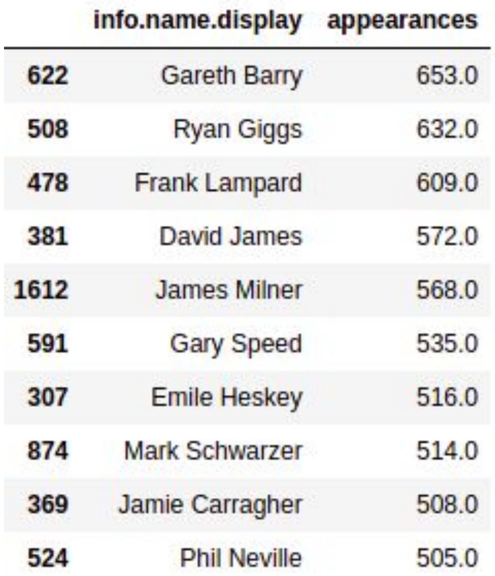
- Player: Discard if 0 appearances, Replace with mean
- Match: Replace with mean

Approach

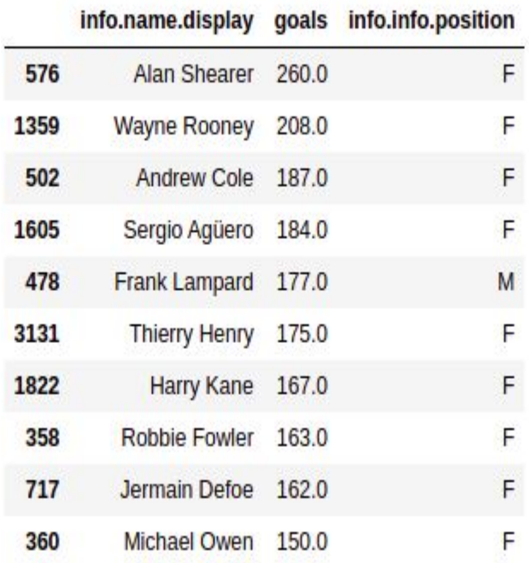
- Player analysis
- Match analysis for season 20/21
- Club analysis for season 20/21
- Specific club all season

Player stats

- > 20 features for players
- 6589 players



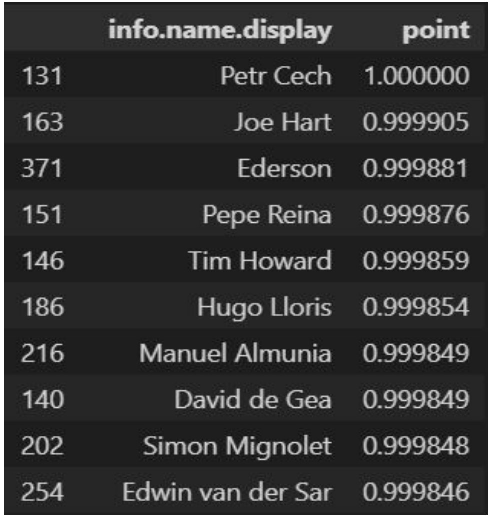
	info.name.display	appearances
622	Gareth Barry	653.0
508	Ryan Giggs	632.0
478	Frank Lampard	609.0
381	David James	572.0
1612	James Milner	568.0
591	Gary Speed	535.0
307	Emile Heskey	516.0
874	Mark Schwarzer	514.0
369	Jamie Carragher	508.0
524	Phil Neville	505.0



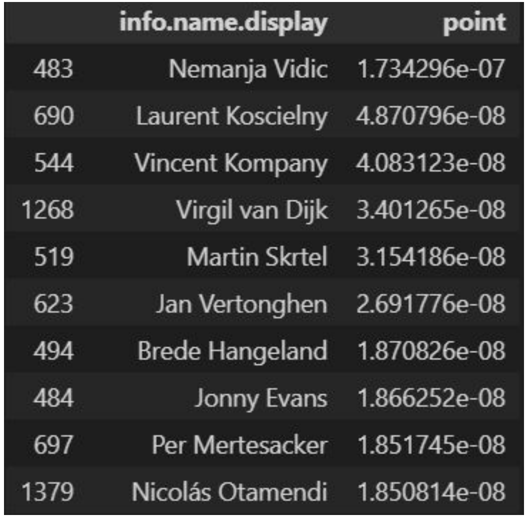
	info.name.display	goals	info.info.position
576	Alan Shearer	260.0	F
1359	Wayne Rooney	208.0	F
502	Andrew Cole	187.0	F
1605	Sergio Agüero	184.0	F
478	Frank Lampard	177.0	M
3131	Thierry Henry	175.0	F
1822	Harry Kane	167.0	F
358	Robbie Fowler	163.0	F
717	Jermain Defoe	162.0	F
360	Michael Owen	150.0	F

Player rank

- Score player using stats
- Position score:
  - GK
  - DF
  - MD
  - FW

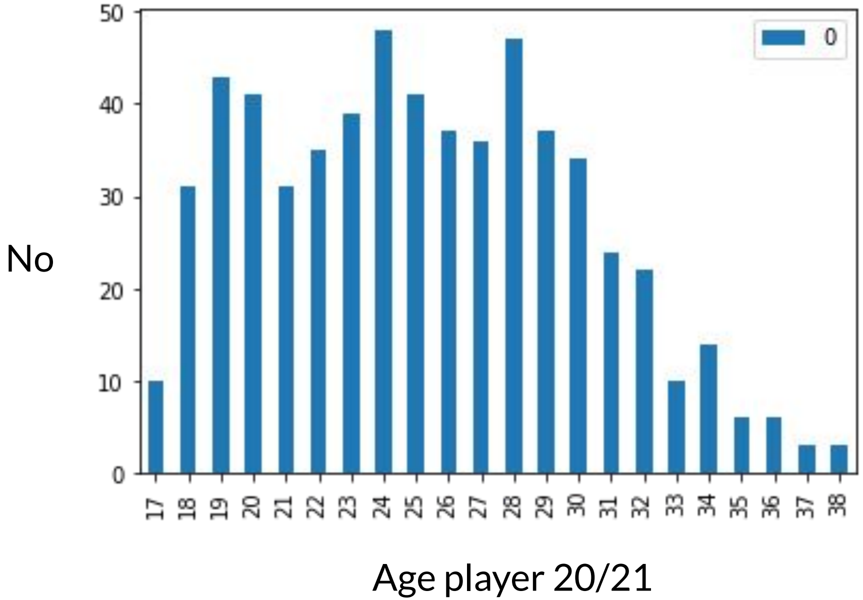


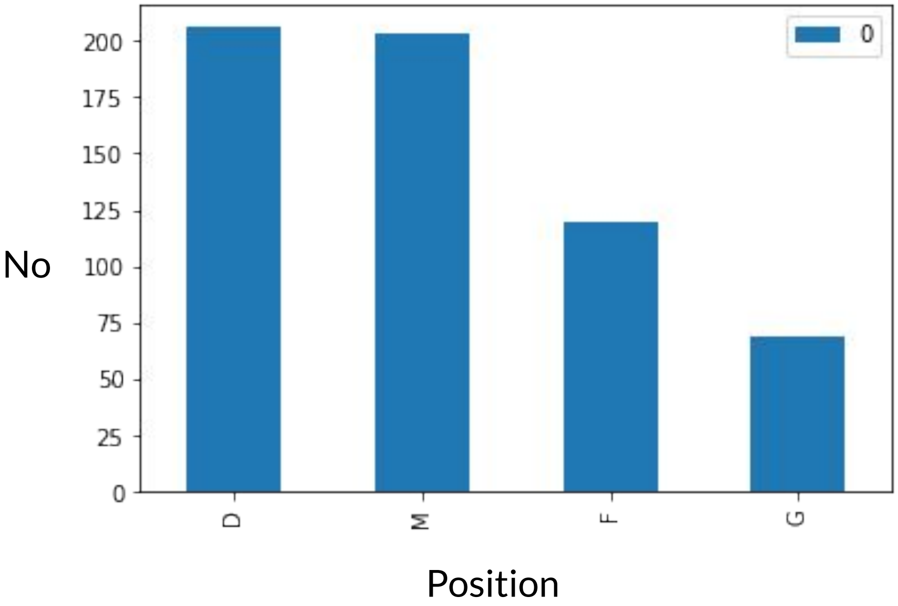
	info.name.display	point
131	Petr Cech	1.000000
163	Joe Hart	0.999905
371	Ederson	0.999881
151	Pepe Reina	0.999876
146	Tim Howard	0.999859
186	Hugo Lloris	0.999854
216	Manuel Almunia	0.999849
140	David de Gea	0.999849
202	Simon Mignolet	0.999848
254	Edwin van der Sar	0.999846



	info.name.display	point
483	Nemanja Vidic	1.734296e-07
690	Laurent Koscielny	4.870796e-08
544	Vincent Kompany	4.083123e-08
1268	Virgil van Dijk	3.401265e-08
519	Martin Skrtel	3.154186e-08
623	Jan Vertonghen	2.691776e-08
494	Brede Hangeland	1.870826e-08
484	Jonny Evans	1.866252e-08
697	Per Mertesacker	1.851745e-08
1379	Nicolás Otamendi	1.850814e-08

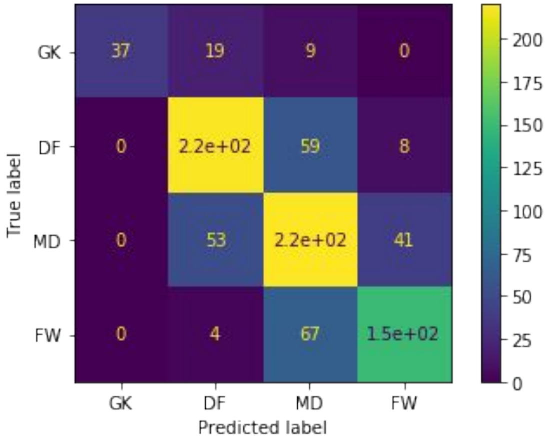
Age and Position distribution





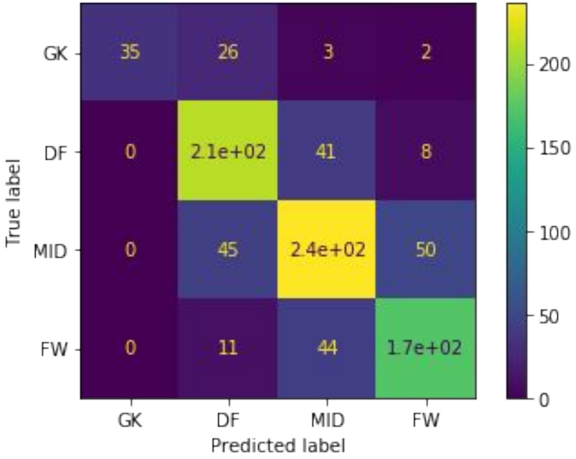
Position prediction

- Input:
  - Player stats: 1 player 1 data point
  - Preprocessing
- Output
  - 4 classes: GK, DF, 'MD, FW
- Simple MLP head vs SVM
- 70% compared with 74%



	GK	DF	MD	FW
GK	37	19	9	0
DF	0	2.2e+02	59	8
MD	0	53	2.2e+02	41
FW	0	4	67	1.5e+02

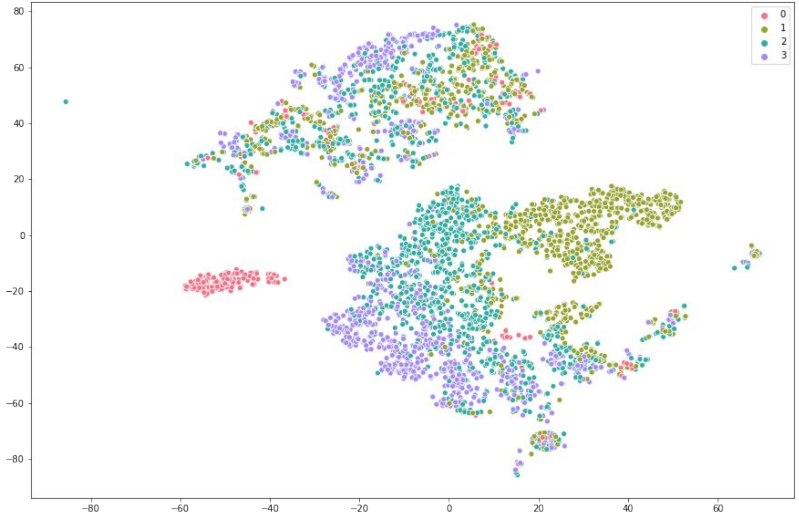
SVM



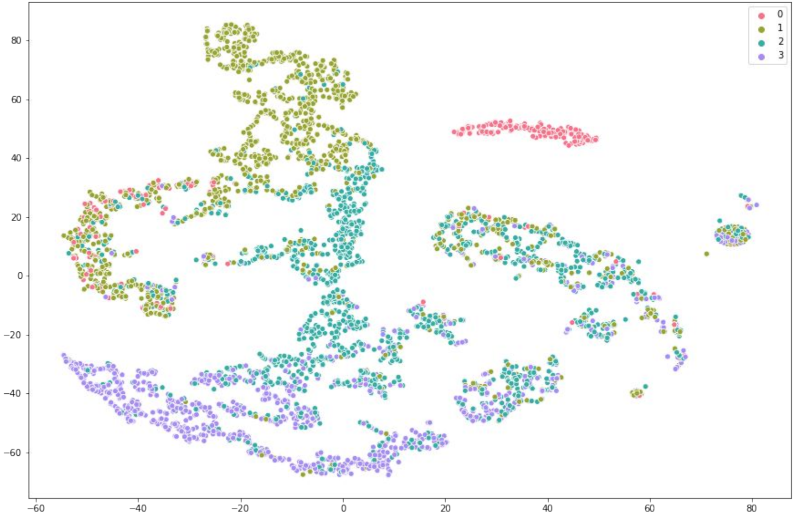
	GK	DF	MID	FW
GK	35	26	3	2
DF	0	2.1e+02	41	8
MID	0	45	2.4e+02	50
FW	0	11	44	1.7e+02

MLP

Tsne for player space



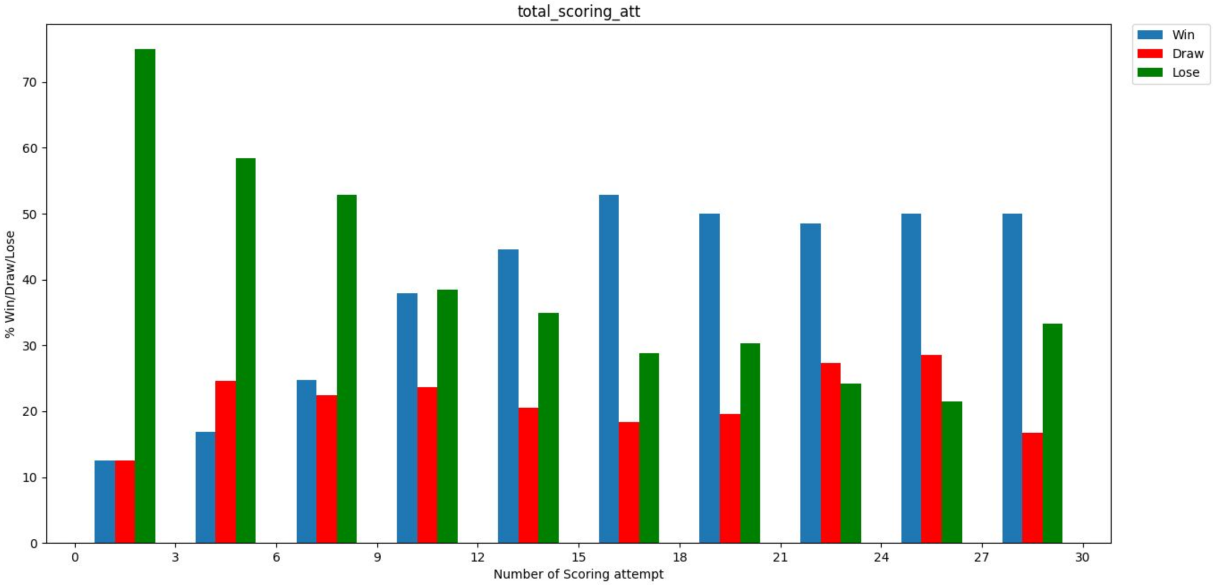
Raw features



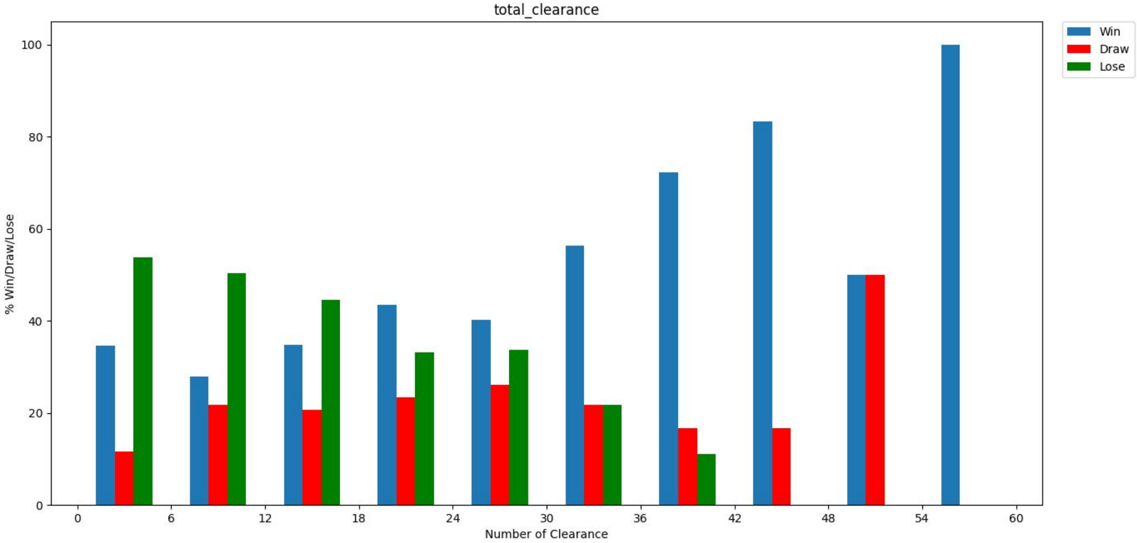
Features after training simple head prediction

Match Stats 20/21

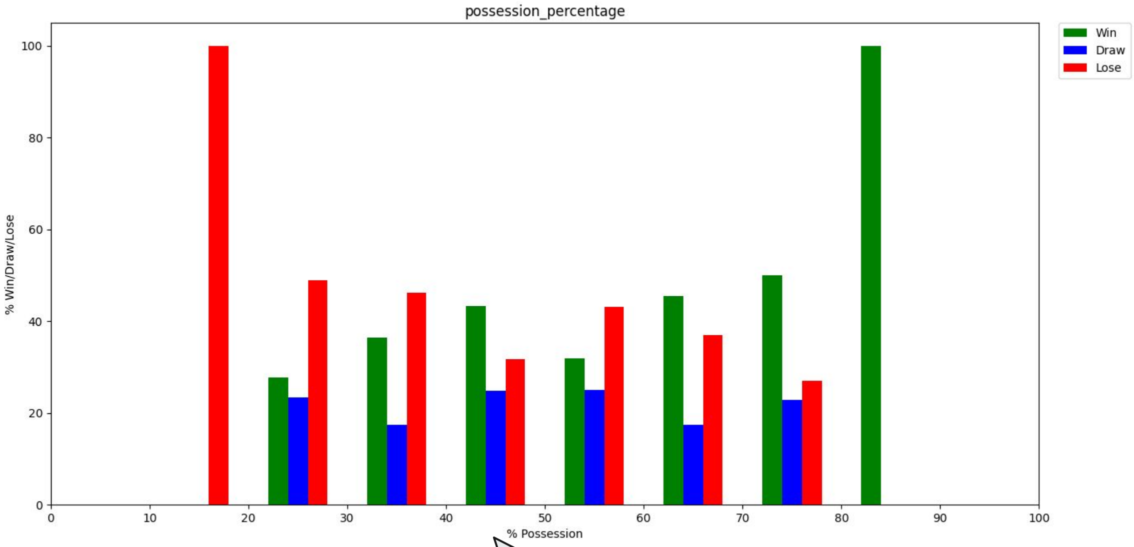
Scoring attempts determine wins



Clearance determine wins

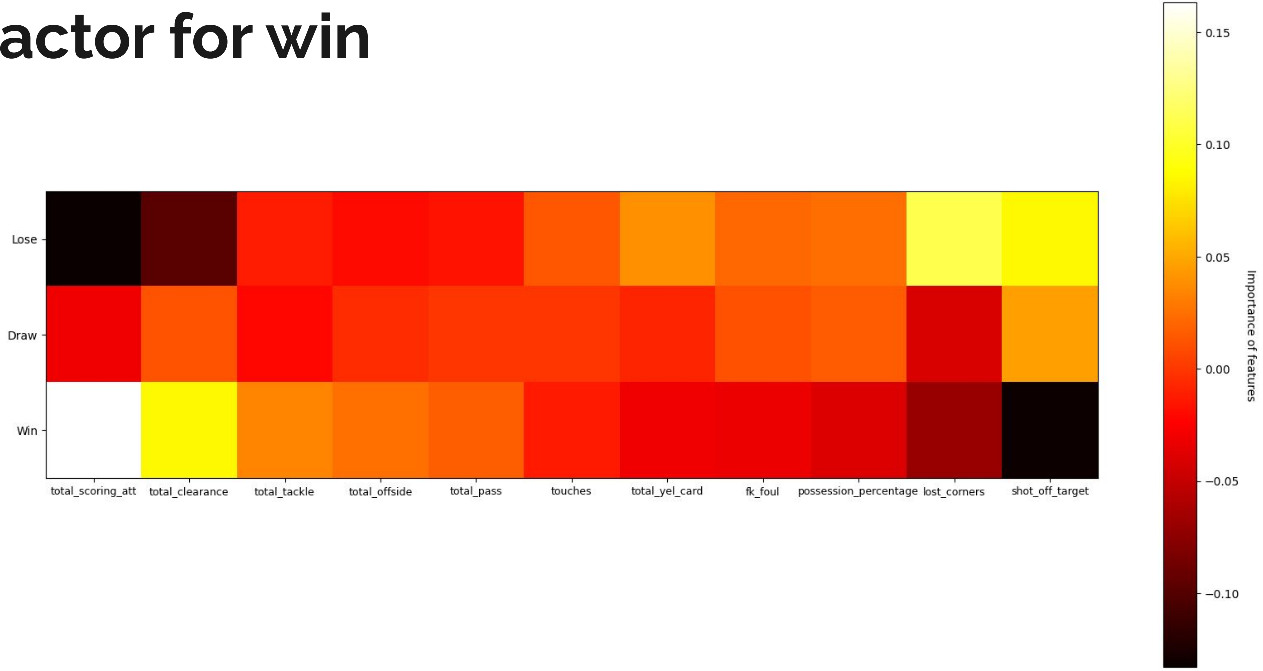


Possession determine wins

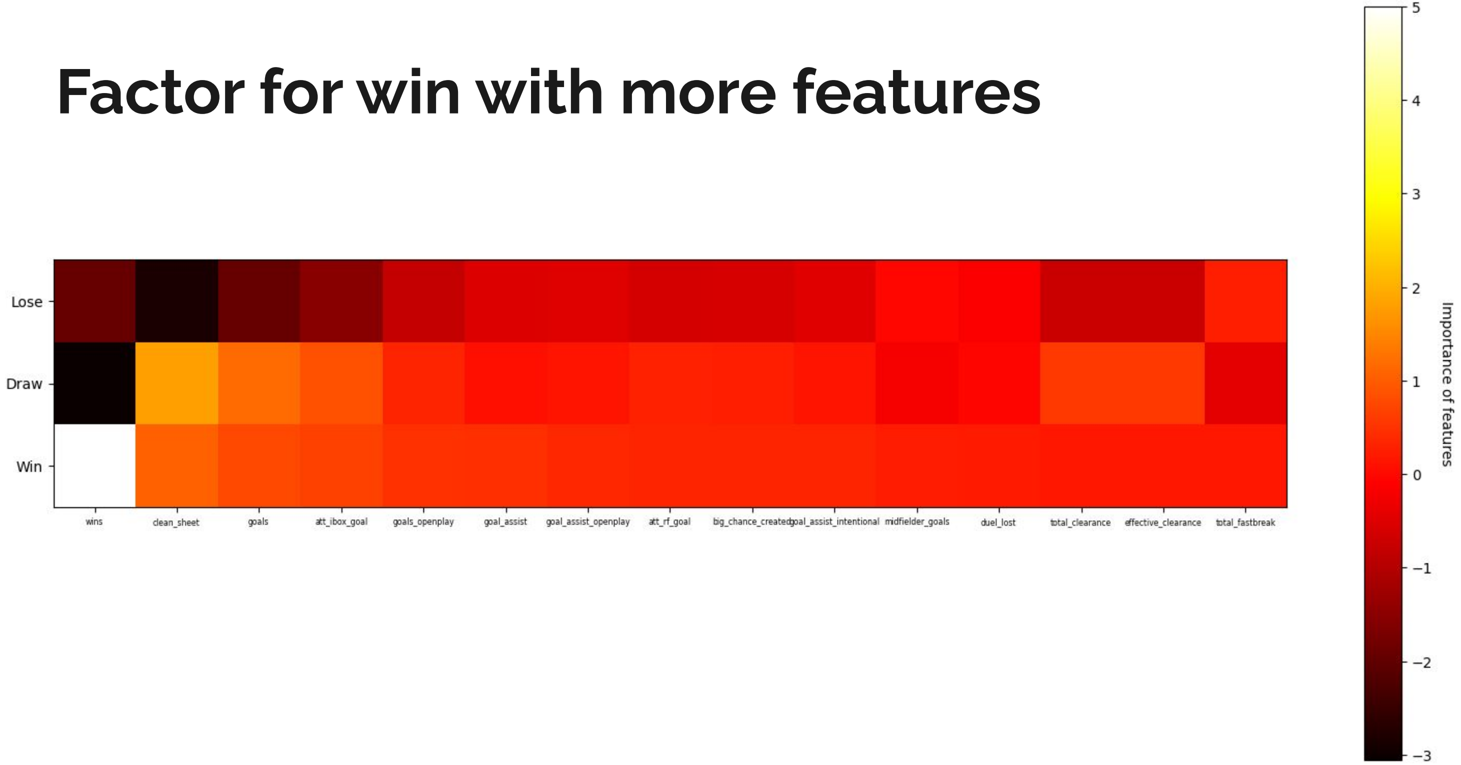


Tweedledum and Tweedledee

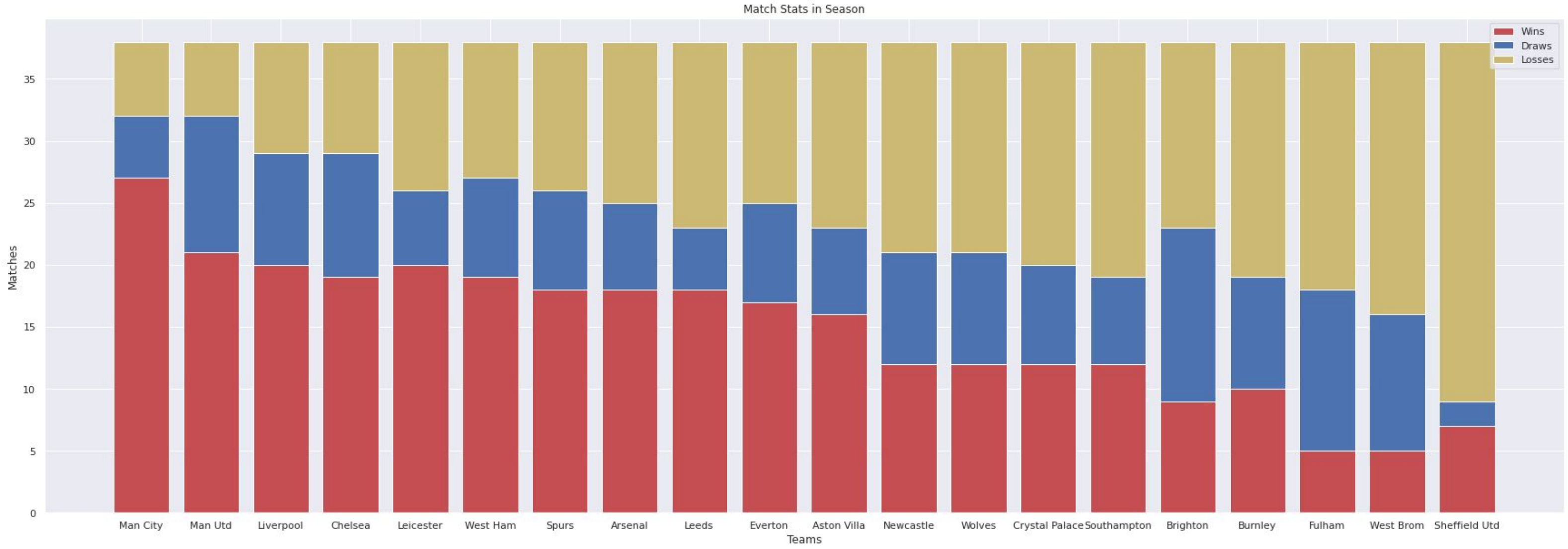
Factor for win



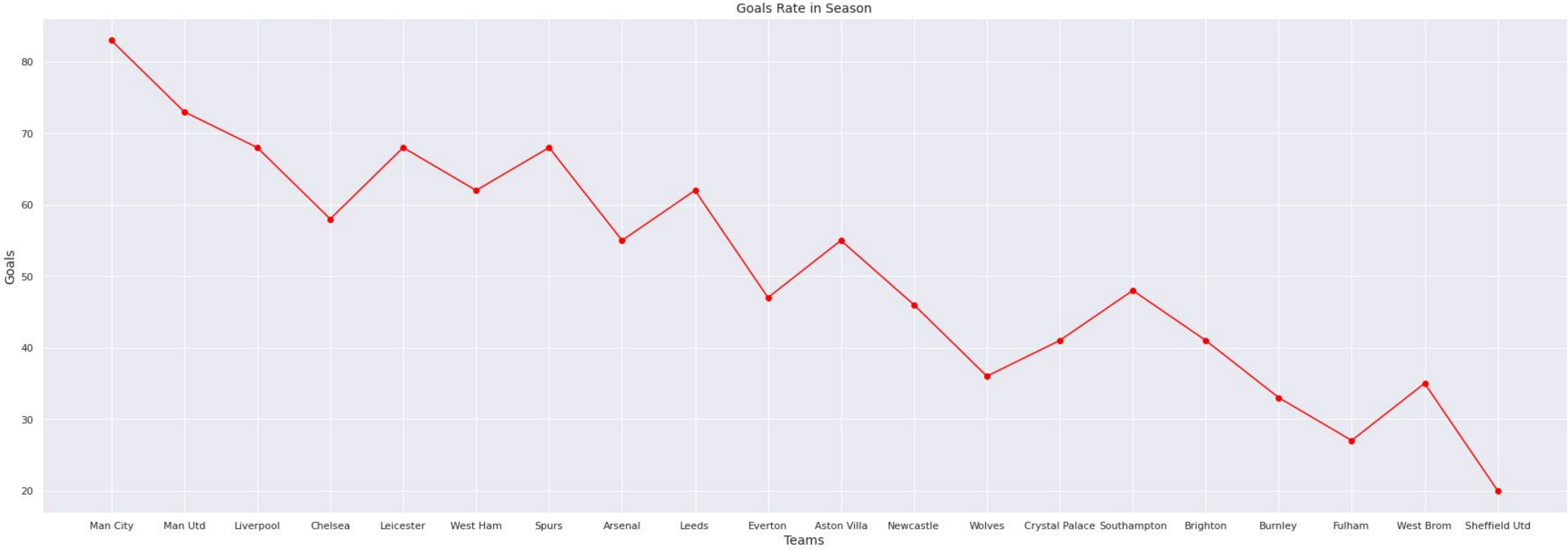
Factor for win with more features



Club Stats in Season 20/21

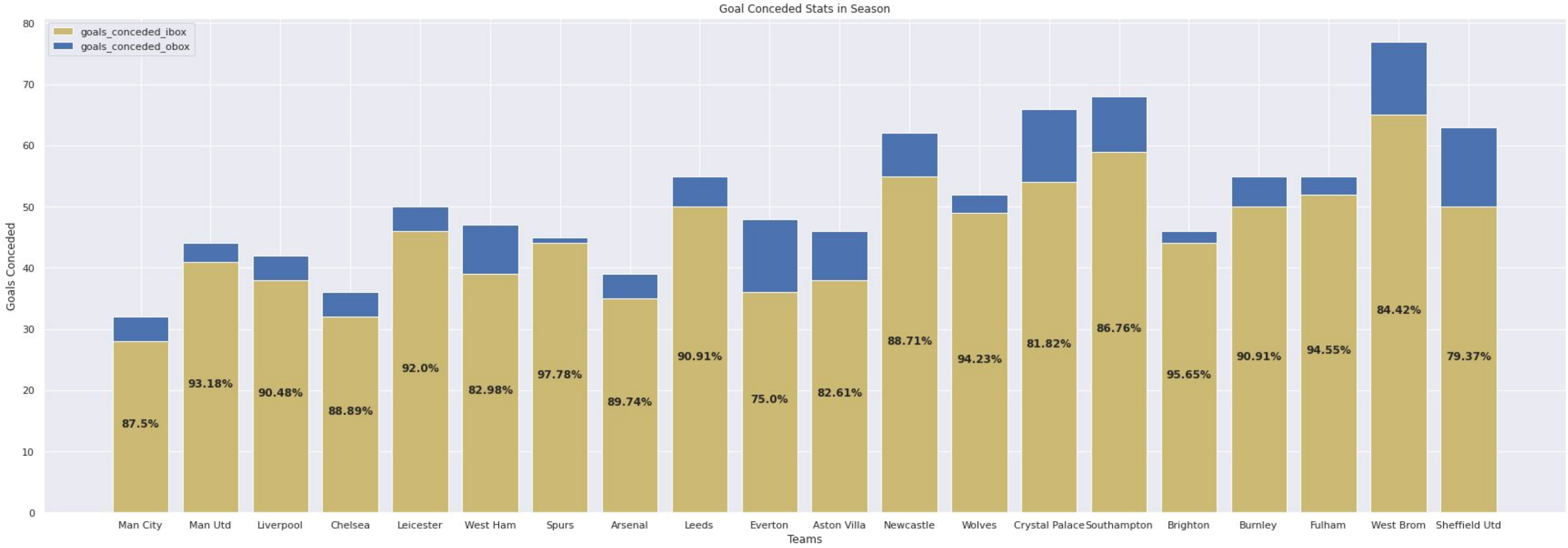


Goals



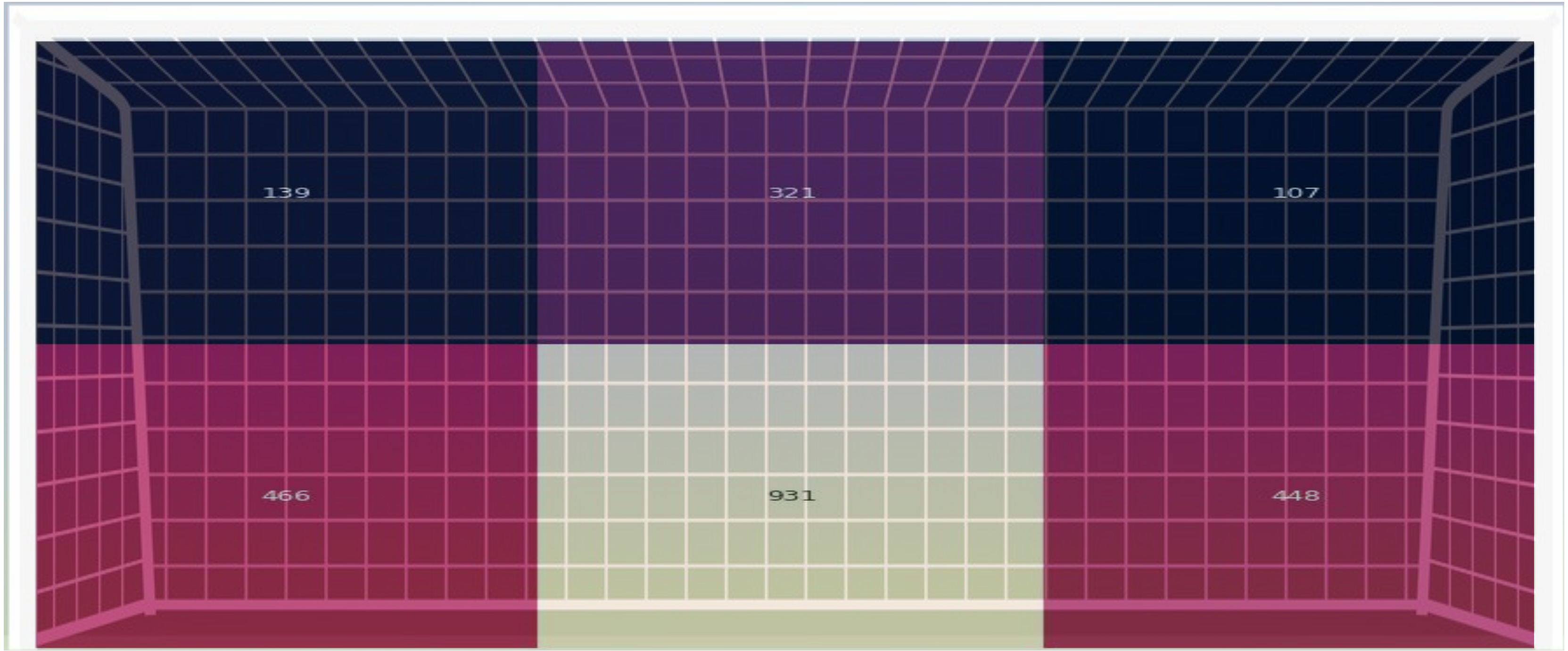
- Number of goals tends to decrease with each team’s position

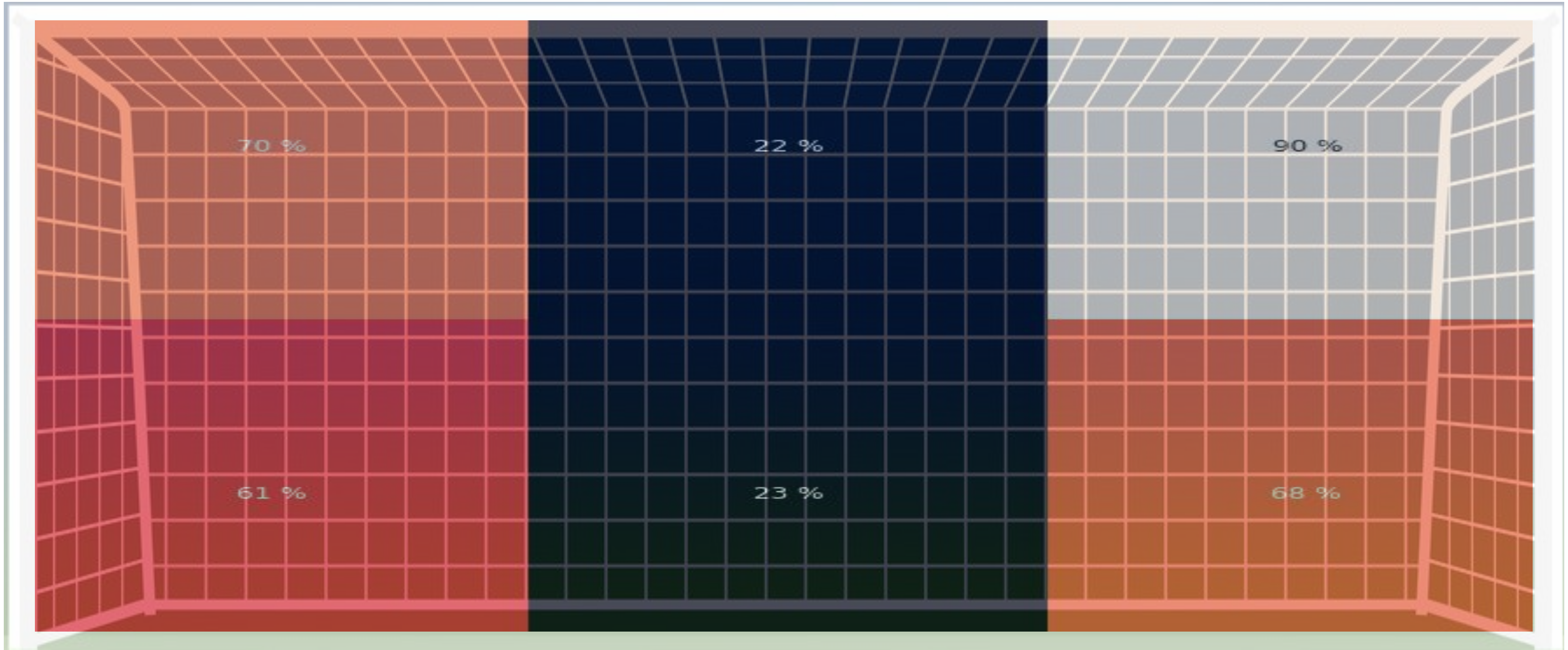
Goal Conceded



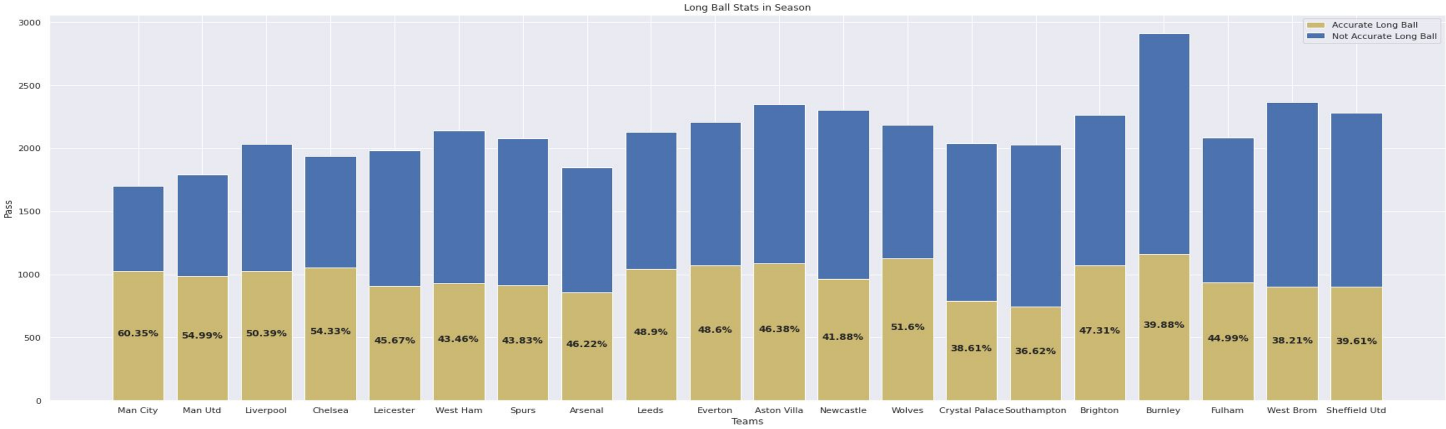
- Spur defenses outbox well
- All clubs conceded more goals inbox

Shot



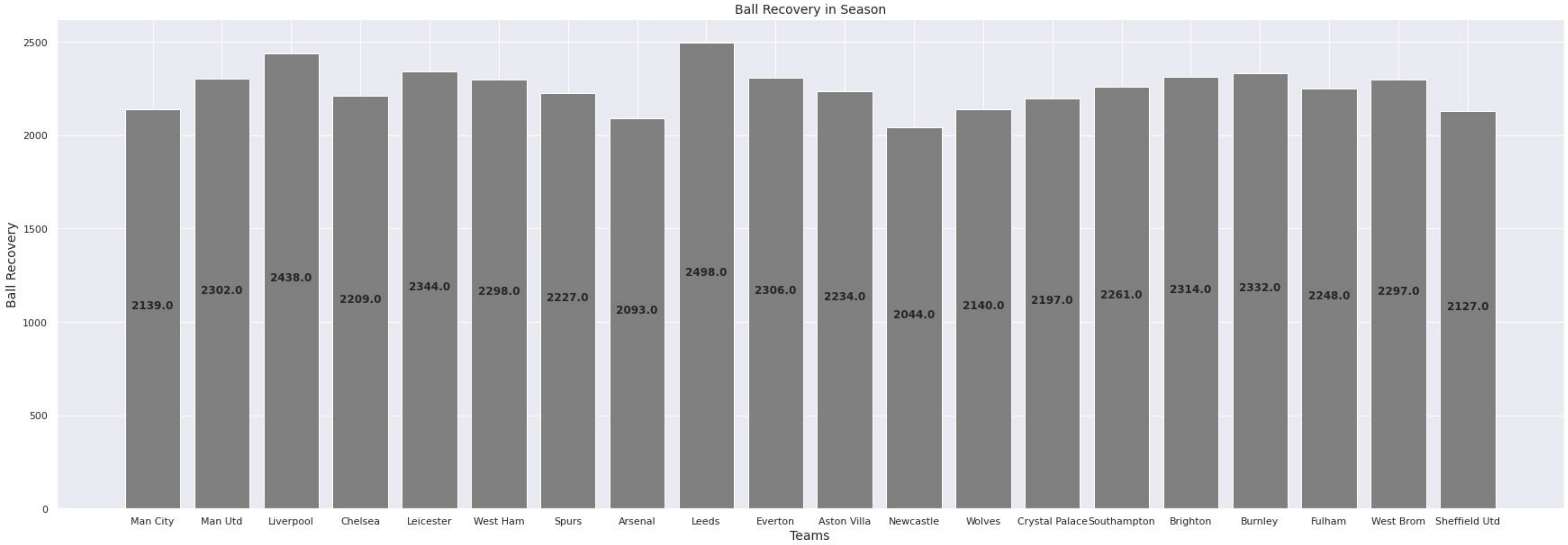

70 %
22 %
90 %
61 %
23 %
68 %

Long balls



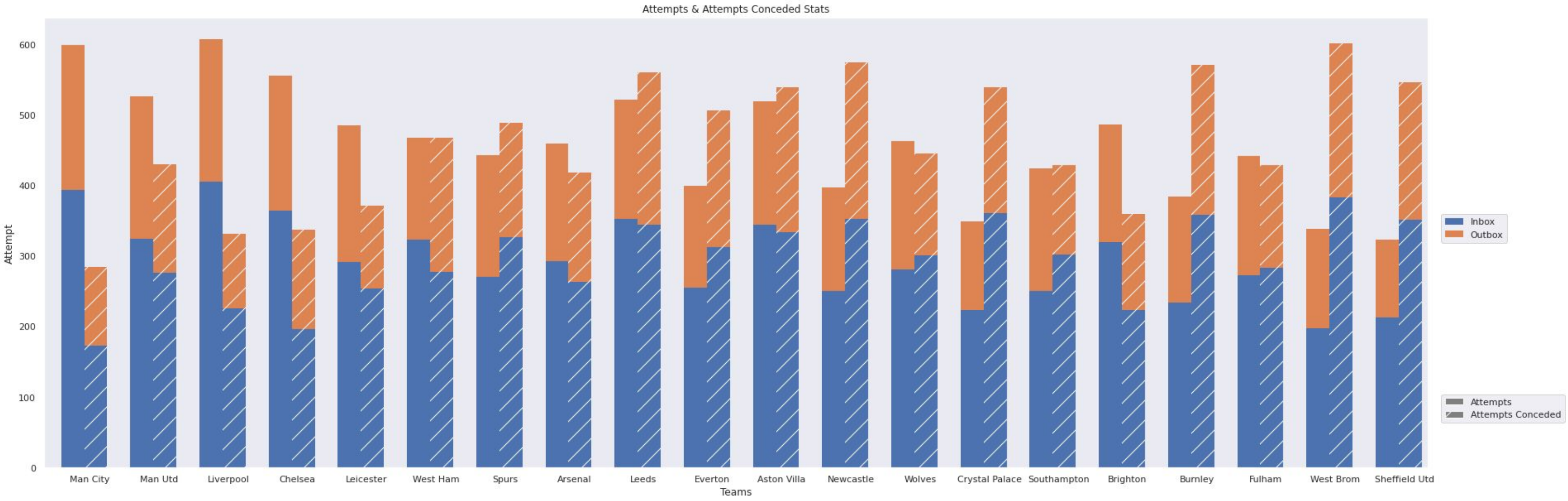
- Fast
- Most used by weak clubs

Pressing



- Liverpool and Leeds have the **pressing** style

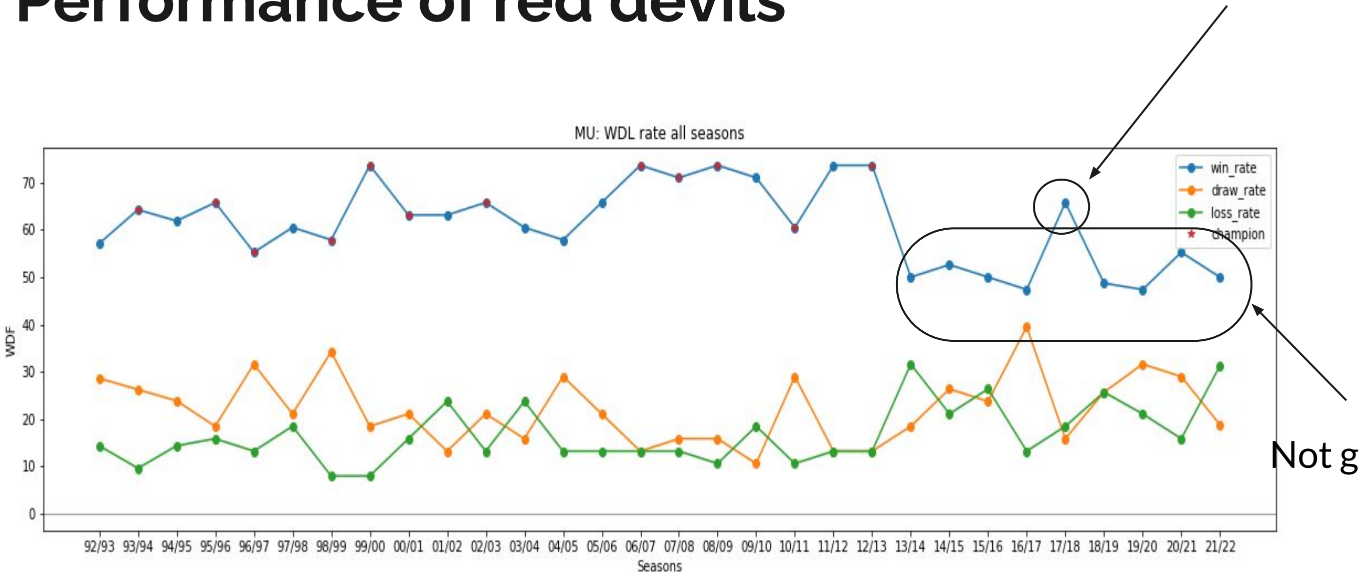
Attempt & Attempt Conceded



- Top teams created more chances than they gave

Manchester United stats

Performance of red devils

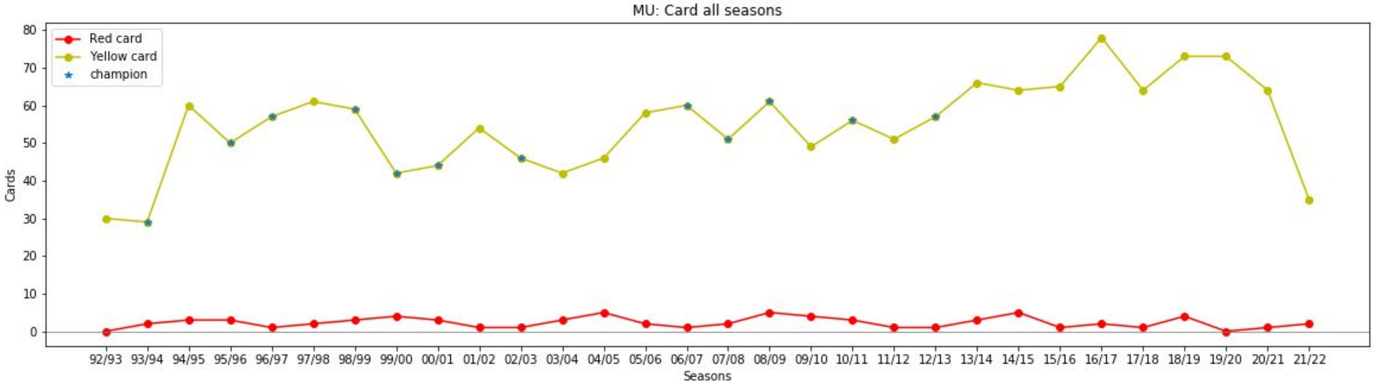


not luck

Not good to be a champ

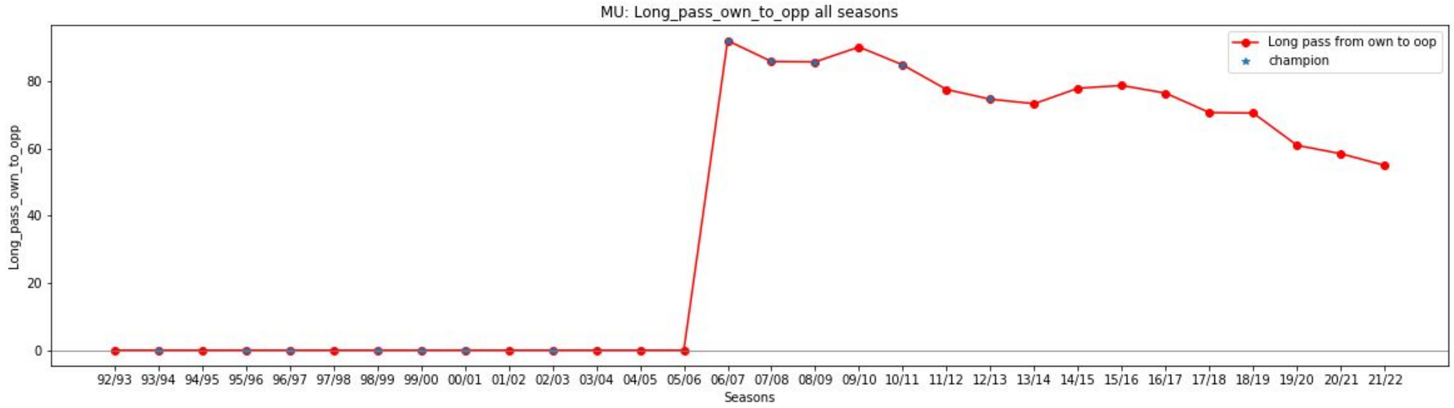
- Worse performance after Sir Alex time

Fouls Card



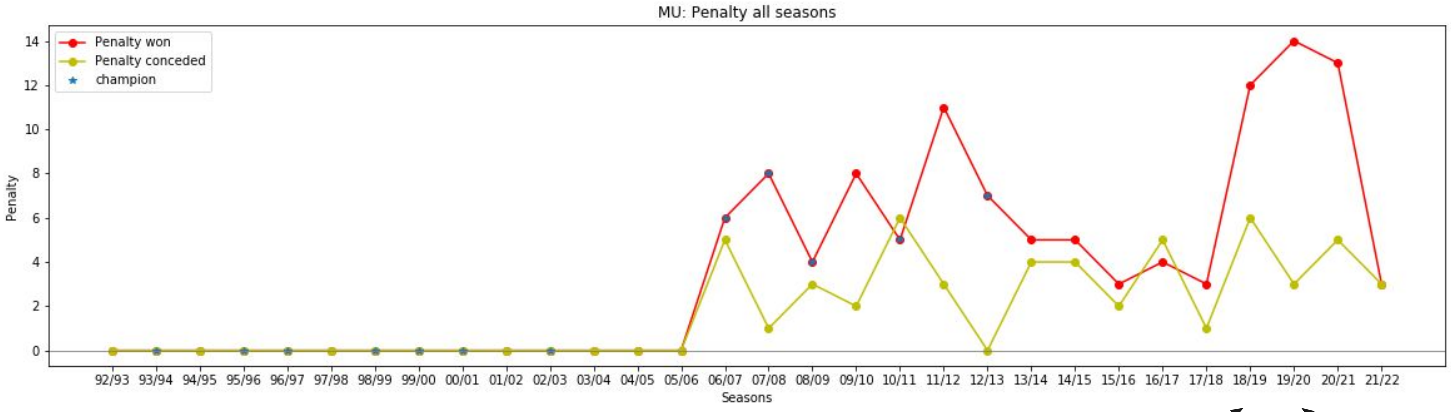
- More aggressive than Sir Alex time

Long pass



- One more times, Long ball still is not the trend

VAR impact

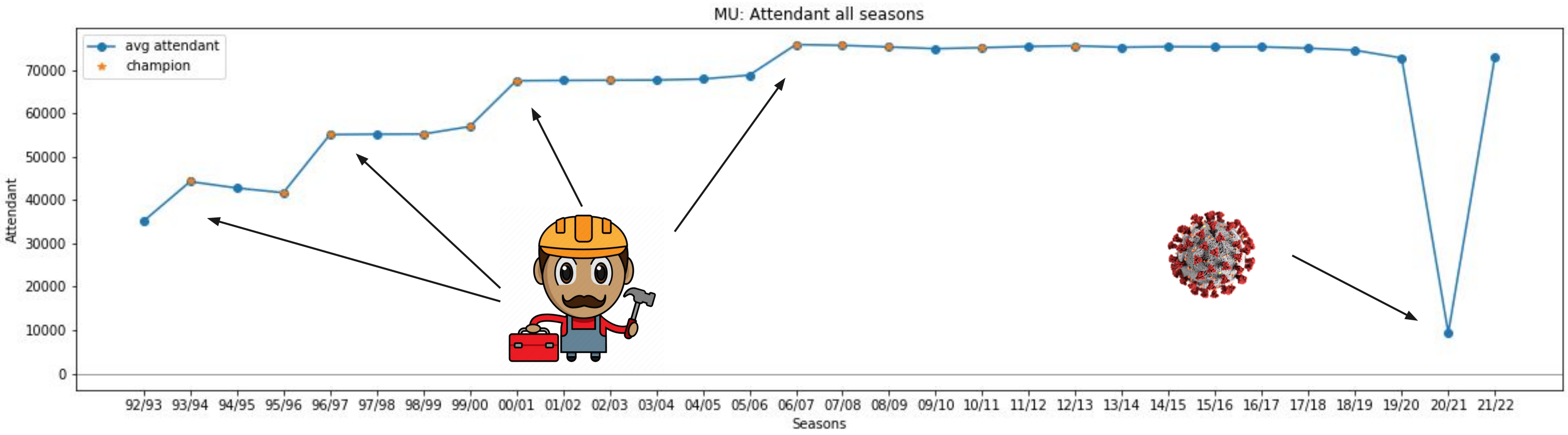


Testing

Official apply

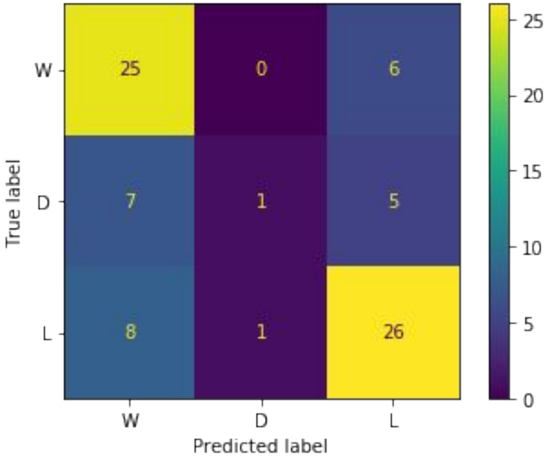
- MU likes VAR ?

Theater of dreams



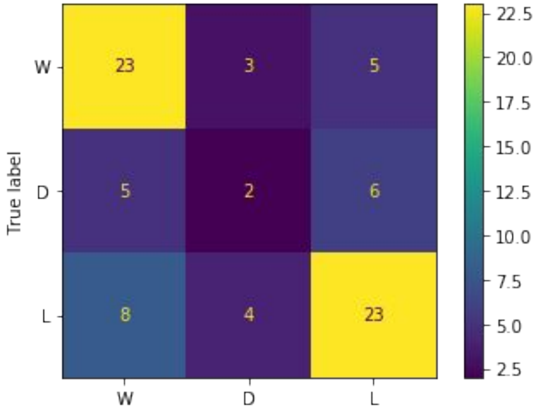
Match prediction

- Using 80% match to predict the rest 20% according time
- Input for predictions:
  - Head to Head
  - Player stats in line up
- Output:
  - Win, Draw or Lose
- Metrics: Accuracy
- Result:
  - Hard to determine **draw**



SVM

65%



NB

60%

Summary

- Get the stats of the EPL teams and the statistics of the seasons.
- Data visualization to comment on statistics for certain seasons and teams
- Use some basic machine learning model to predict and evaluate.

Thanks for listening