ltapp.yatinkapur.me

Capturing team image and performances over time using game states.

Purpose

- Get away from mainstream metrics like wins, shots, passes, to interpret – or misinterpret meaning of performances
- Use how dominant teams are with respect to sustaining dominant periods of game states
  - There are three game states
    - Winning
    - Level (draw)
    - Losing

Metrics Used for Analysis

- Time spent leading
- Time spent trailing
- Goals scored
- Goals conceded

Key Positions

- Positions 1, 2, 3, 4

this is good

- Positions 5, 6, 7*

¯_(ツ)_/¯

- Positions 18, 19, 20

this is bad

Matches:
- pk: match_id

```
mysql> describe matches;
+-------------+-------------+------+-----+---------+-------+
| Field       | Type        | Null | Key | Default | Extra |
+-------------+-------------+------+-----+---------+-------+
| match_id    | int(11)     | NO   | PRI | NULL    |       |
| home_team   | varchar(45) | YES  |     | NULL    |       |
| away_team   | varchar(45) | YES  |     | NULL    |       |
| competition | varchar(45) | YES  |     | NULL    |       |
| date        | varchar(45) | YES  |     | NULL    |       |
| matchday    | int(11)     | NO   | PRI | NULL    |       |
+-------------+-------------+------+-----+---------+-------+
6 rows in set (0.05 sec)
```

```
mysql> describe scores;
+------------+---------+------+-----+---------+-------+
| Field      | Type    | Null | Key | Default | Extra |
+------------+---------+------+-----+---------+-------+
| match_id   | int(11) | NO   | PRI | NULL    |       |
| minute     | int(11) | NO   | PRI | NULL    |       |
| home_score | int(11) | NO   | PRI | NULL    |       |
| away_score | int(11) | NO   | PRI | NULL    |       |
+------------+---------+------+-----+---------+-------+
4 rows in set (0.05 sec)
```

Scores:
- pk: match_id, minute, home_score, away_score
- stores the minute of the game where the score changed, used to extrapolate data into extended_scores table

Extended_Scores:

- pk: match_id, minute

- stores record of every game at every minute and the score at that minute

```
mysql> describe extended_scores;
+------------+---------+------+-----+---------+-------+
| Field      | Type    | Null | Key | Default | Extra |
+------------+---------+------+-----+---------+-------+
| match_id   | int(11) | NO   | PRI | NULL    |       |
| home_score | int(11) | NO   |     | NULL    |       |
| away_score | int(11) | NO   |     | NULL    |       |
| minute     | int(11) | NO   | PRI | NULL    |       |
+------------+---------+------+-----+---------+-------+
4 rows in set (0.06 sec)
```

```
mysql> describe competition_summary;
+----------------+-------------+------+-----+---------+-------+
| Field          | Type        | Null | Key | Default | Extra |
+----------------+-------------+------+-----+---------+-------+
| competition    | varchar(45) | NO   | PRI | NULL    |       |
| team           | varchar(45) | NO   | PRI | NULL    |       |
| pos            | int(11)     | NO   |     | NULL    |       |
| pts            | int(11)     | YES  |     | NULL    |       |
| gp             | int(11)     | YES  |     | NULL    |       |
| gs             | int(11)     | YES  |     | NULL    |       |
| ga             | int(11)     | YES  |     | NULL    |       |
| gd             | int(11)     | YES  |     | NULL    |       |
| lead_time      | int(11)     | YES  |     | NULL    |       |
| trail_time     | int(11)     | YES  |     | NULL    |       |
| lead_time_p90  | int(11)     | YES  |     | NULL    |       |
| trail_time_p90 | int(11)     | YES  |     | NULL    |       |
| top_four       | float       | YES  |     | NULL    |       |
| relegation     | float       | YES  |     | NULL    |       |
| top_six        | float       | YES  |     | NULL    |       |
+----------------+-------------+------+-----+---------+-------+
15 rows in set (0.05 sec)
```

Competition_Summary:

- pk: competition, team

- stores summary data for every team in every season, including predictions for finishing positions

Data Collection Process

- Data collected from https://www.football-lineups.com/tourn/FA_Premier_League_2018-2019/
- Each match page has records for:
  - Match ID (primary key for most tables)
  - Home and Away Teams
  - Goal Times (stored in scores table)
- Use beautifulsoup, MySQLdb, requests_html, numpy, pandas libraries to help with data collection and feature engineering

File Structure

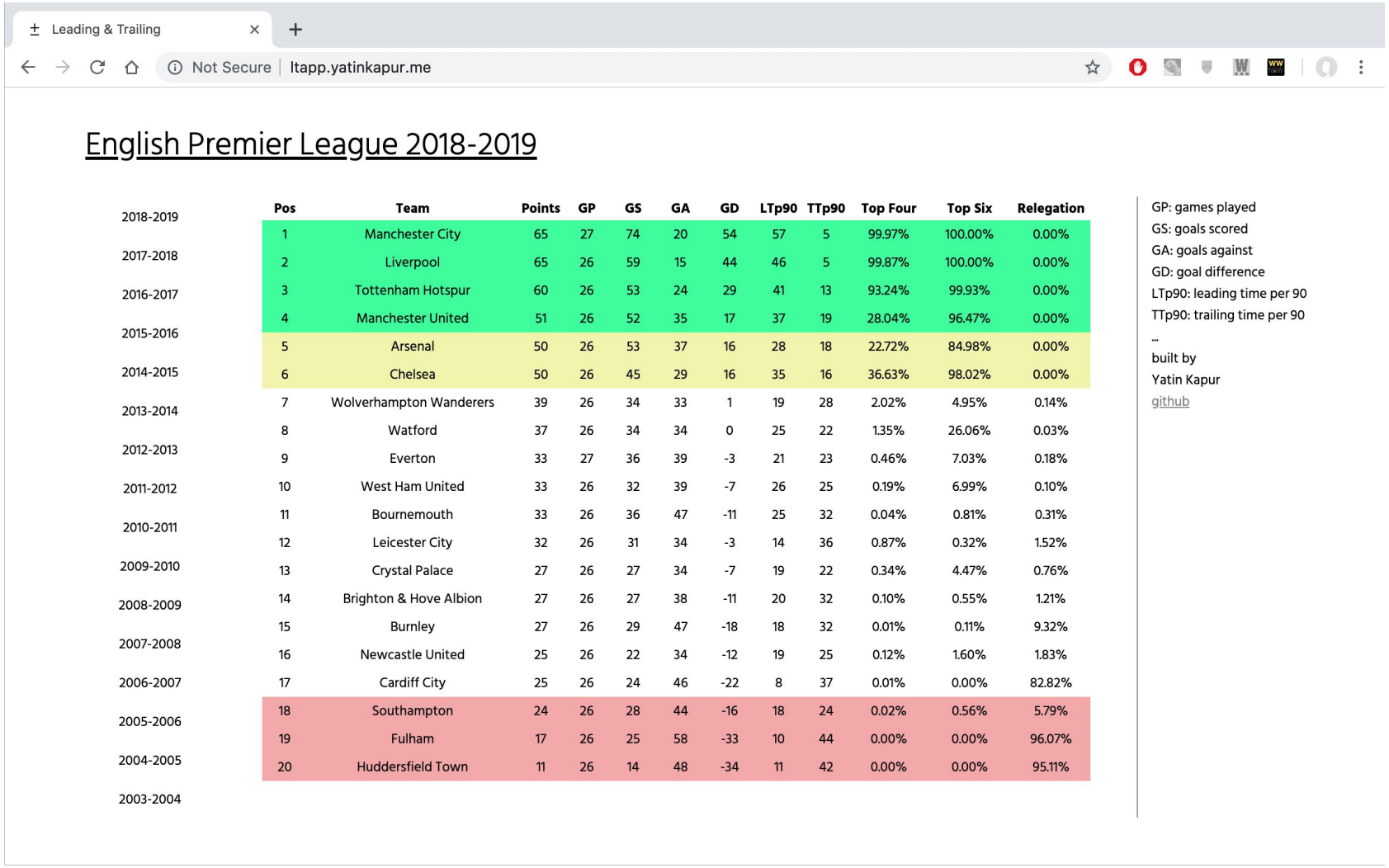
```
.
├── Pipfile
├── Pipfile.lock
├── README.md
├── __init__.py
├── add_game_entry.py
├── config.ini
├── create_leading.py
├── dbconfig.py
├── insert.py
├── metadate.py
├── model.py
├── russia_leading.py
├── update_matches.py
├── update_meta.py
└── update_standings.py
```

- **update_matches.py** loads in all the matches that are not already loaded into the db
- **add_game_entry.py** adds the specific game entry, and the score timings into scores table
- **create_leading.py** uses scores data to populate minute by minute scores into extended_scores table
- **update_standings.py** calculates time spent trailing & leading, for teams based on extended_scores data
  - also calculates probability of finishing in certain positions using **model.py** (logistic regression model)

Delivery

Fetch data from competition_summary to display on homepage:

- frontend: D3.js
- backend: flask, ajax calls



English Premier League 2018-2019

2018-2019
2017-2018
2016-2017
2015-2016
2014-2015
2013-2014
2012-2013
2011-2012
2010-2011
2009-2010
2008-2009
2007-2008
2006-2007
2005-2006
2004-2005
2003-2004

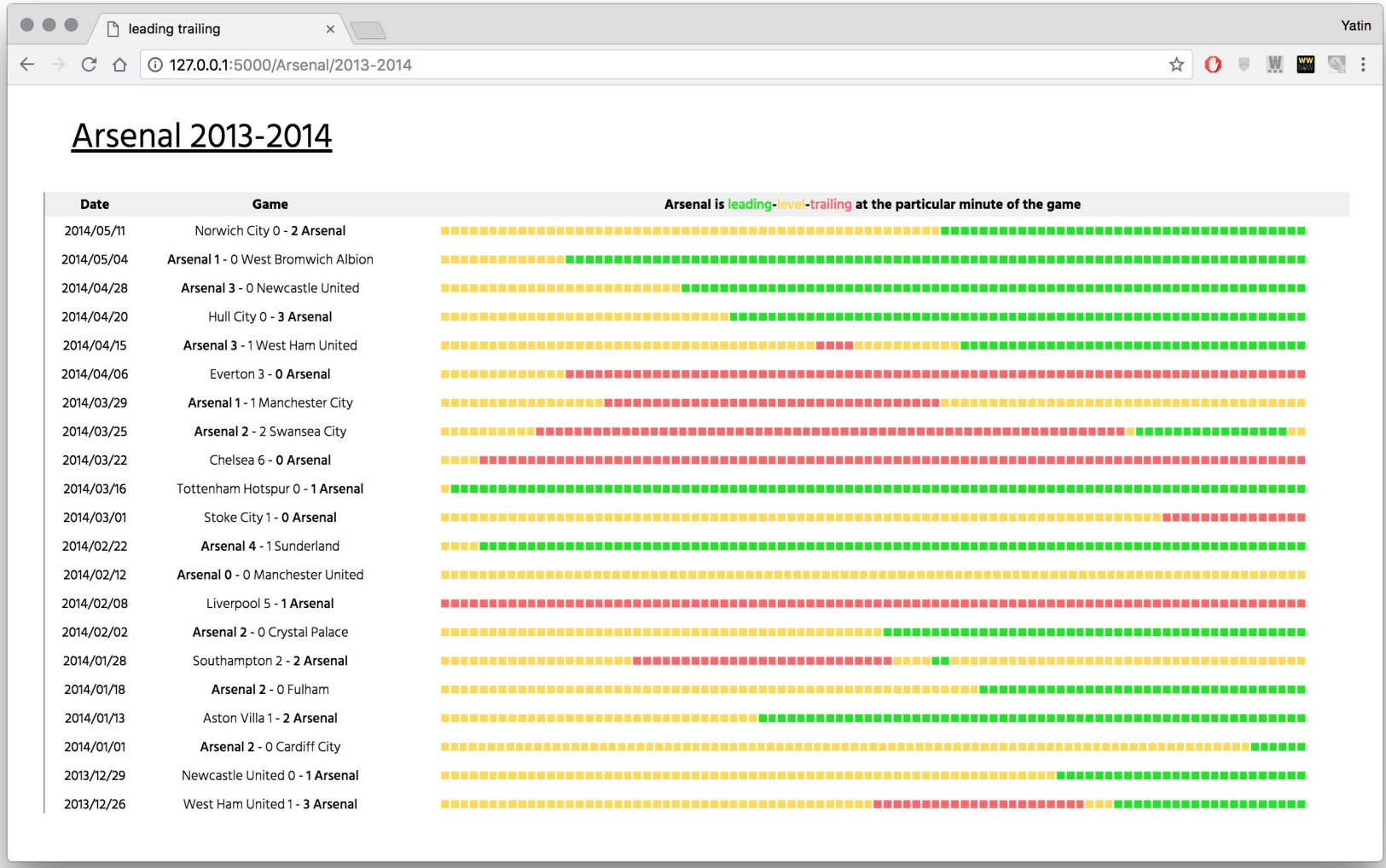
Pos	Team	Points	GP	GS	GA	GD	LTp90	TTp90	Top Four	Top Six	Relegation
1	Manchester City	65	27	74	20	54	57	5	99.97%	100.00%	0.00%
2	Liverpool	65	26	59	15	44	46	5	99.87%	100.00%	0.00%
3	Tottenham Hotspur	60	26	53	24	29	41	13	93.24%	99.93%	0.00%
4	Manchester United	51	26	52	35	17	37	19	28.04%	96.47%	0.00%
5	Arsenal	50	26	53	37	16	28	18	22.72%	84.98%	0.00%
6	Chelsea	50	26	45	29	16	35	16	36.63%	98.02%	0.00%
7	Wolverhampton Wanderers	39	26	34	33	1	19	28	2.02%	4.95%	0.14%
8	Watford	37	26	34	34	0	25	22	1.35%	26.06%	0.03%
9	Everton	33	27	36	39	-3	21	23	0.46%	7.03%	0.18%
10	West Ham United	33	26	32	39	-7	26	25	0.19%	6.99%	0.10%
11	Bournemouth	33	26	36	47	-11	25	32	0.04%	0.81%	0.31%
12	Leicester City	32	26	31	34	-3	14	36	0.87%	0.32%	1.52%
13	Crystal Palace	27	26	27	34	-7	19	22	0.34%	4.47%	0.76%
14	Brighton & Hove Albion	27	26	27	38	-11	20	32	0.10%	0.55%	1.21%
15	Burnley	27	26	29	47	-18	18	32	0.01%	0.11%	9.32%
16	Newcastle United	25	26	22	34	-12	19	25	0.12%	1.60%	1.83%
17	Cardiff City	25	26	24	46	-22	8	37	0.01%	0.00%	82.82%
18	Southampton	24	26	28	44	-16	18	24	0.02%	0.56%	5.79%
19	Fulham	17	26	25	58	-33	10	44	0.00%	0.00%	96.07%
20	Huddersfield Town	11	26	14	48	-34	11	42	0.00%	0.00%	95.11%

GP: games played
GS: goals scored
GA: goals against
GD: goal difference
LTp90: leading time per 90
TTp90: trailing time per 90
...
built by
Yatin Kapur
github

Team Summaries

Highlight trends for the game states of a team, how frequently they are there, what patterns are there, stretches of wins/losses, etc.

- data from extended_scores
- visualization: D3.js



Model(s) – Logistic regression to determine probability of achieving a final position in one of the target groups

- Training Data: 1997-2018 Seasons data

```
mysql> select * from competition_summary limit 5;
```

competition	team	pos	pts	gp	gs	ga	gd	lead_time	trail_time	lead_time_p90	trail_time_p90	top_four	top_six	relegation
FA_Premier_League_1997-1998	Arsenal	1	78	38	68	33	35	1423	352	37	9	1	1	0
FA_Premier_League_1997-1998	Aston Villa	7	57	38	49	48	1	831	996	22	26	0	0	0
FA_Premier_League_1997-1998	Barnsley	19	35	38	37	82	-45	808	1209	21	32	0	0	1
FA_Premier_League_1997-1998	Blackburn Rovers	6	58	38	57	52	5	1056	955	28	25	0	1	0
FA_Premier_League_1997-1998	Bolton Wanderers	18	40	38	41	61	-20	539	1109	14	29	0	0	1

5 rows in set (0.04 sec)

- Features: [leading_time_p90, trailing_time_p90, gs, ga]
- Target: one of {top_four, relegation, top_six}

Is the model any good?

Pos	Team	Points	GP	GS	GA	GD	LTp90	TTp90	Top Four	Top Six	Relegation
1	Manchester City	65	27	74	20	54	57	5	99.97%	100.00%	0.00%
2	Liverpool	65	26	59	15	44	46	5	99.87%	100.00%	0.00%
3	Tottenham Hotspur	60	26	53	24	29	41	13	93.24%	99.93%	0.00%
4	Manchester United	51	26	52	35	17	37	19	28.04%	96.47%	0.00%
5	Arsenal	50	26	53	37	16	28	18	22.72%	84.98%	0.00%
6	Chelsea	50	26	45	29	16	35	16	36.63%	98.02%	0.00%

TEAM	SPI	OFF.	DEF.	W	D	L	GOAL DIFF.	PTS.	RELEGATED	QUALIFY FOR UCL	WIN PREMIER LEAGUE
Liverpool 65 pts	92.1	2.8	0.3	28.4	7.2	2.4	+62	92	–	>99%	50%
Man. City 65 pts	93.9	3.1	0.3	29.4	3.6	5.0	+76	92	–	>99%	48%
Tottenham 60 pts	85.0	2.5	0.5	26.3	2.6	9.1	+37	82	–	96%	2%
Chelsea 50 pts	83.7	2.4	0.5	21.5	7.7	8.8	+25	72	–	46%	<1%
Arsenal 50 pts	78.7	2.3	0.7	20.8	7.9	9.3	+22	70	–	31%	<1%
Man. United 51 pts	78.9	2.3	0.7	20.3	8.9	8.8	+21	70	–	28%	<1%

Is it?!

15	Burnley	27	26	29	47	-18	18	32	0.01%	0.11%	9.32%
16	Newcastle United	25	26	22	34	-12	19	25	0.12%	1.60%	1.83%
17	Cardiff City	25	26	24	46	-22	8	37	0.01%	0.00%	82.82%
18	Southampton	24	26	28	44	-16	18	24	0.02%	0.56%	5.79%
19	Fulham	17	26	25	58	-33	10	44	0.00%	0.00%	96.07%
20	Huddersfield Town	11	26	14	48	-34	11	42	0.00%	0.00%	95.11%

Burnley 27 pts	62.4	1.7	1.0	10.0	8.8	19.2	-26	39	18%	<1%	–
Southampton 24 pts	65.5	1.8	0.9	8.7	12.1	17.2	-20	38	21%	<1%	–
Cardiff City 25 pts	58.9	1.6	1.0	9.8	7.0	21.2	-32	36	38%	<1%	–
Fulham 17 pts	58.2	1.7	1.2	6.8	7.9	23.3	-42	28	91%	<1%	–
Huddersfield 11 pts	56.6	1.4	0.9	4.2	8.1	25.6	-44	21	>99%	–	–

Improvements

- Could use histograms for team summaries
- Date snapshots for a team/league over a period of time
- Add different leagues
- Add sliders to show changes over time