CRICKET

PREDICTING AN INDIVIDUAL'S LIKELIHOOD OF MAKING SENIOR NATIONAL TEAM AFTER REPRESENTING COUNTRY AT U-19 (YOUTH) LEVEL

WHY & IMPLICATIONS

- **I represented NY U-19, never selected for US U-19**
- **What metrics are representative of senior national team selection?**
  - Measurable? Politics? Scout's gut feeling?
- **Scouts can run a similar algorithm to predict a player's success at the senior national level**

WHAT'S CRICKET ANYWAY?

- **Team 1 scores as many runs as possible (1st inning)**
- **Team 2 tries to score the target that was set (2nd inning)**
- **11 players on a team**
  - Batsmen
    - Score as many runs as possible
  - Bowlers (Pitchers)
    - Prevent batsmen from scoring runs ; get them out AKA wicket
    - 2 main types (spinners and fast bowlers)
  - Wicketkeepers (Catcher)
    - Usually very good batsman

THE DATA

- **Statsguru on cricinfo.com**
- **~4000 U-19 players -- matching their names against ~1600 senior national players since 1970s**

In [213]:
```
senior.head()
```

Out[213]:

	name	tenure	MP	total_runs	hs	avg_runs	centuries	halfcenturies	best_bowling	avg_bowling	5W	catchings	stumping
0	SR Tendulkar (India)	1989-2012	463	18426	200*	44.83	49	154	1932-05-01	44.48	2	140	0
1	DPMD Jayawardene (Asia/SL)	1998-2015	448	12650	144	33.37	19	8	1956-02-01	70.37	0	218	0
2	ST Jayasuriya (Asia/SL)	1989-2011	445	13430	189	32.36	28	323	2016-06-29	36.75	4	123	0
3	KC Sangakkara (Asia/ICC/SL)	2000-2015	404	14234	169	41.98	25	NaN	NaT	NaN	NaN	402	99
4	Shahid Afridi (Asia/ICC/Pak)	1996-2015	398	8064	124	23.57	6	395	2016-07-12	34.51	9	127	0

In [212]:
```
youth.head()
```

Out[212]:

	name	tenure	MP	total_runs	hs	avg_runs	centuries	halfcenturies	best_bowling	avg_bowling	5W	catchings	stumpings	avg_dif
0	Nazmul Hossain Shanto (BD19)	2013-2016	58	1820	113*	37.91	2	13	1948-03-01 00:00:00	20.23	0	24	0	17.68
1	Mahmudul Hasan (BD19)	2007-2010	57	1168	82*	23.36	0	66	2016-04-17 00:00:00	22.19	0	33	0	1.16
2	Mehedi Hasan Miraz (BD19)	2013-2016	56	1305	87	29.00	0	80	2016-05-17 00:00:00	20.90	1	20	0	8.10
3	Imad Wasim (Pak19)	2005-2008	49	638	85	26.58	0	73	1938-05-01 00:00:00	21.19	1	16	0	5.39
4	Joyraz Sheik (BD19)	2013-2016	43	1130	90	28.97	0	NaN	NaN	NaN	NaN	14	0	NaN

WHAT'S LEFT?

- **I need to match 'youth' names with 'senior' names.**
  - If youth name is in 'senior' dataset 1 else 0→ 'labels' column (y)
- **Break out country from name feature to its own**
- **Tenure field split into 'start career' & 'end career'**
- **'Best bowling' feat came in as date. The 'month' is # of wickets. I want to get that figure alone for each player.**
- **Manipulating various object datatypes**
- **Feat engineer: how many 50s/100s scored out of matches played? How many wickets per match?**

BIGGEST PROBLEMS

- What if the youth player will be selected but hasn’t been yet? (they are currently still playing U-19)

BEST PREDICTOR PREDICTION

- Combination of Clustering and Logistic Regression or RF
    - Predicting 3 or 4 clusters (batsmen, bowlers, and WKs)
    - Classification

- No free lunch