CRICKET

PREDICTING WHETHER OR NOT AN INDIVIDUAL WILL MAKE SENIOR NATIONAL TEAM AFTER REPRESENTING COUNTRY AT U-19 (YOUTH) LEVEL

WHAT'S CRICKET AGAIN?

- **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
    - 2 main types (spinners and fast bowlers)

WHY & IMPLICATIONS

- What metrics are representative of senior national team selection?

Measurable? Politics? Scout's gut feeling?

- **Business Application: Scouting**

- **Goal: Who will make it, who won't?**

- **Priorities**
  - What features determine "making it"
  - Prediction accuracy & Interpretability

THE DATA

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

youth['made_it'] = youth.name1.isin(senior['name1']).astype(int)

17% of all players “make_it”

	dum_avg_score	metric
0	0.503	precision
1	0.177	recall
2	0.262	f1
3	0.503	accuracy

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

SINCE THEN?

- Matched ‘youth’ names with ‘senior’ names.
- Broke out country from name feature to its own
- Tenure field split into ‘start career’ & ‘end career’
- Feat engineer: newsworthiness/news_fqy

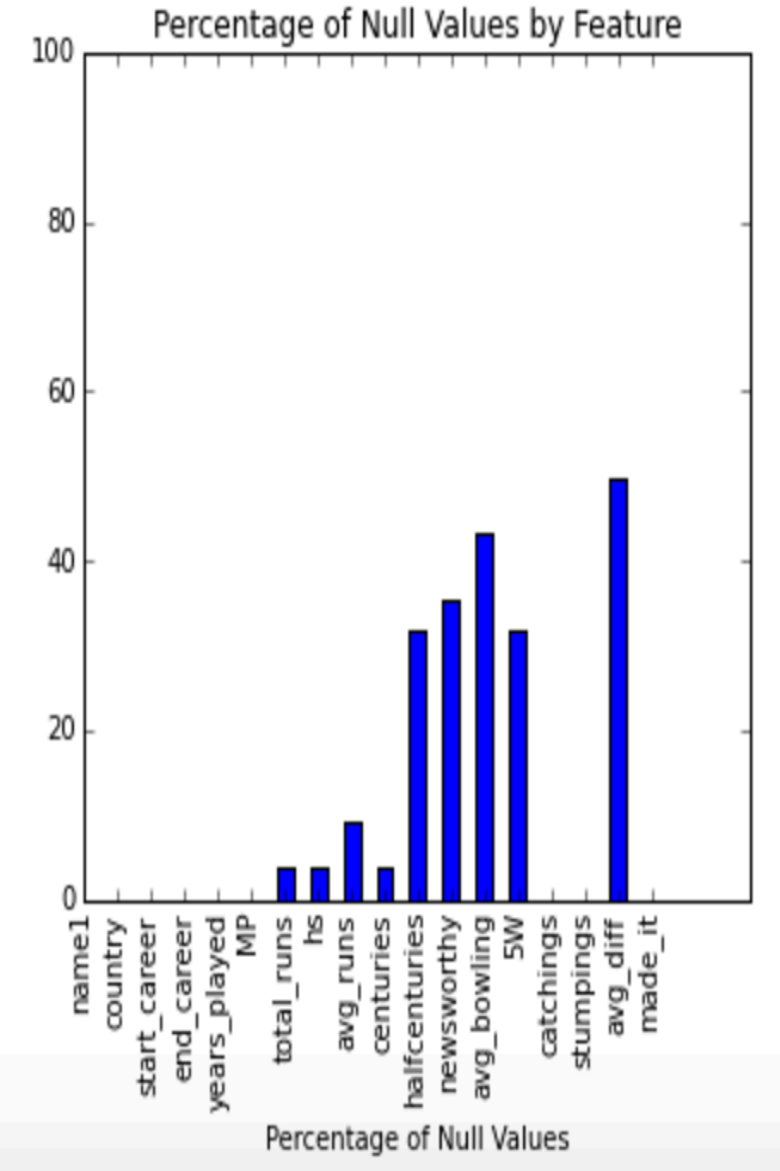
```
In [212]: youth.head()
```

Out[212]:

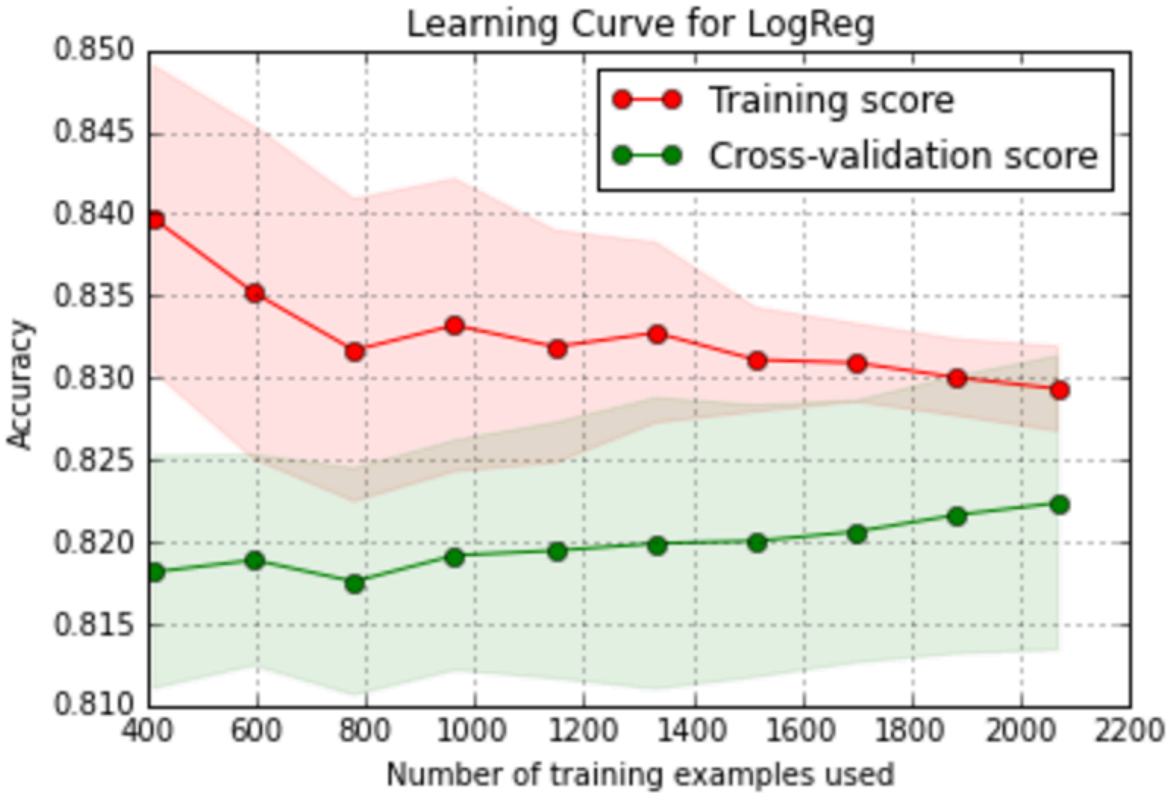
	name	tenure	MP	total_runs	hs	avg_runs	centuries	halfcenturies	best_bowling	avg_bowling	5W	catchings	stumpings	avg_dif
0	Nazmul Hossain Shanto (BD19)	2013-2016	58	1820	113*	37.91	2	13	1948-03-01 00:00:00	20.23	0	24	0	17.68
1	Mahmudul Hasan (BD19)	2007-2010	57	1168	82*	23.36	0	66	2016-04-17 00:00:00	22.19	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

DECISIONS

- **Made all null values zeros**
- **Decided against drop/mean**
  - Most players specialized
- **Clustering to solve specializations**
  - K-means groupings
  - Append label to training set



LOGREG MODEL



Accuracy performance slightly increasing with sample size --variance

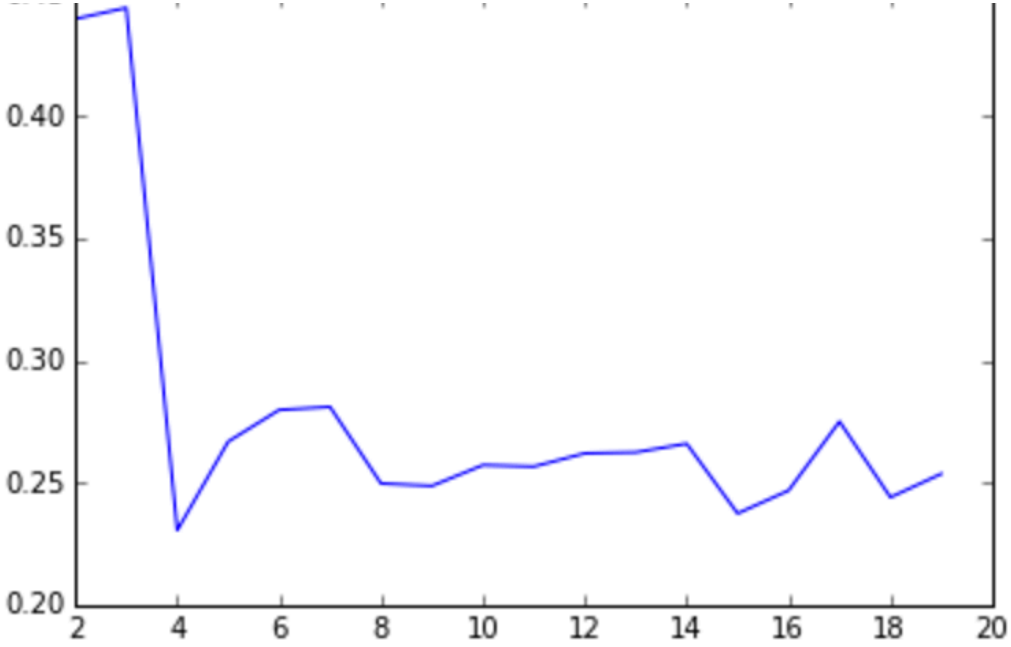
82% of the time the model predicted that the player made it, or didn't make it correctly

MODEL RESULTS

K=2

	dum_avg_score	metric
0	0.503	precision
1	0.177	recall
2	0.262	f1
3	0.503	accuracy

	LR_avg_score	metric
0	0.024	precision
1	0.571	recall
2	0.047	f1
3	0.822	accuracy



CLUSTERS AND COEFFICIENTS

	0	1
years_played	1.528	0.237
MP	15.343	4.420
newsworthy	10.258	2.533
total_runs	312.413	48.469
hs	72.605	23.622
avg_runs	30.041	13.645
centuries	0.312	0.002
halfcenturies	9.917	2.599
avg_bowling	19.391	18.741
5W	0.148	0.018
catchings	6.773	1.425
stumpings	0.620	0.097
avg_diff	-1.415	-10.441

- A lot of weight is put on major performances
- Averages are lesser indicators

Nutshell:

If a player performs well 1 game, you're in better shape than if you consistently performed above average.

	coefs	features
4	0.325	hs
7	0.249	halfcenturies
3	0.229	total_runs
8	0.141	avg_bowling
0	0.136	years_played
11	0.092	stumpings
5	0.078	avg_runs
12	0.060	avg_diff
9	0.024	5W
2	0.000	newsworthy
10	-0.041	catchings
6	-0.058	centuries
1	-0.359	MP

BIGGEST PROBLEMS

- **What if the youth player will be selected but hasn't been yet?**
- **Data cleaning**
  - Converting datatypes
  - Feature engineering implementation
- **Missing data**
- **Different metrics?**

NEXT STEPS

- **Time series (address the issue that some of these players are current)**

```
In [1441]: youth.groupby(by=['end_career', 'made_it'], axis=0).count()
```

2010.000	**0**	174	174	174	174	174	171	171	168	171	126	123	96
	1	41	41	41	41	41	41	41	39	41	33	33	31
2011.000	**0**	64	64	64	64	64	62	62	57	62	51	49	38
2012.000	**0**	202	202	202	202	202	198	198	193	198	149	146	130
	1	20	20	20	20	20	20	20	18	20	15	15	13
2013.000	**0**	65	65	65	65	65	57	57	50	57	52	44	36
2014.000	**0**	233	233	233	233	233	233	233	223	233	173	173	145
	1	9	9	9	9	9	9	9	9	9	8	8	8
2015.000	**0**	77	77	77	77	77	73	73	70	73	49	45	41
2016.000	**0**	246	246	246	246	246	242	242	234	242	174	171	151
	1	2	2	2	2	2	2	2	2	2	1	1	1

- **Add clustering results as a feature and rerun LR**

QUESTIONS?