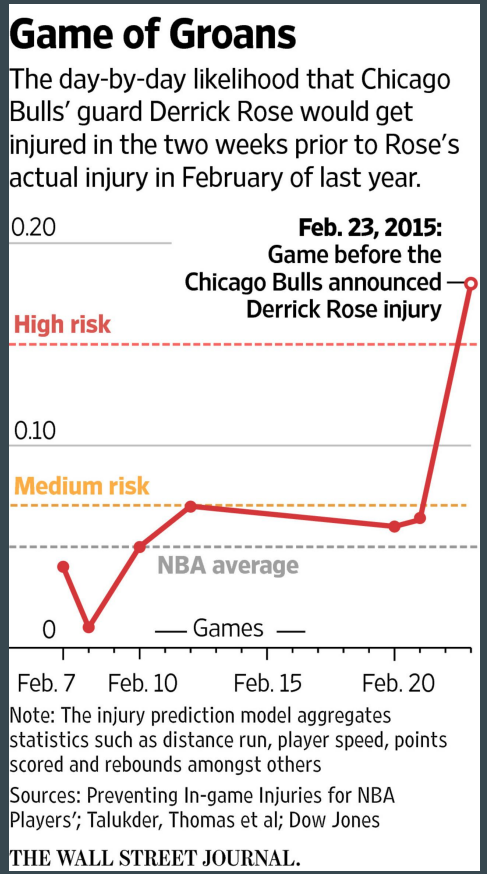
Predicting and Preventing Injuries for NBA Players

Jay Messina
https://github.com/jaymessina3/Honors



- "By resting the top 20% of high risk scores at any given day there is a potential to prevent 60% of all injuries."

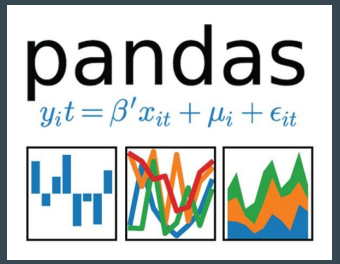
- 2016 MIT Sloan Sports Analytics Conference

Talukder, H., Vincent, T.: Preventing in-game injuries for NBA players (2016)

Agenda

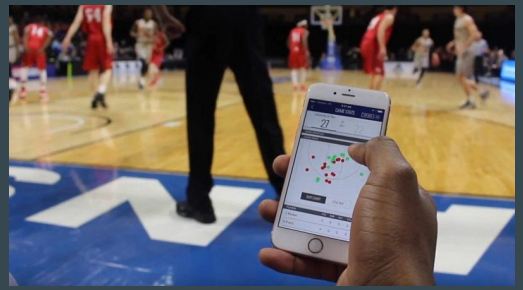
Data Aggregation

Data Organization



Machine Learning

Real Time Model



SportsVu Data

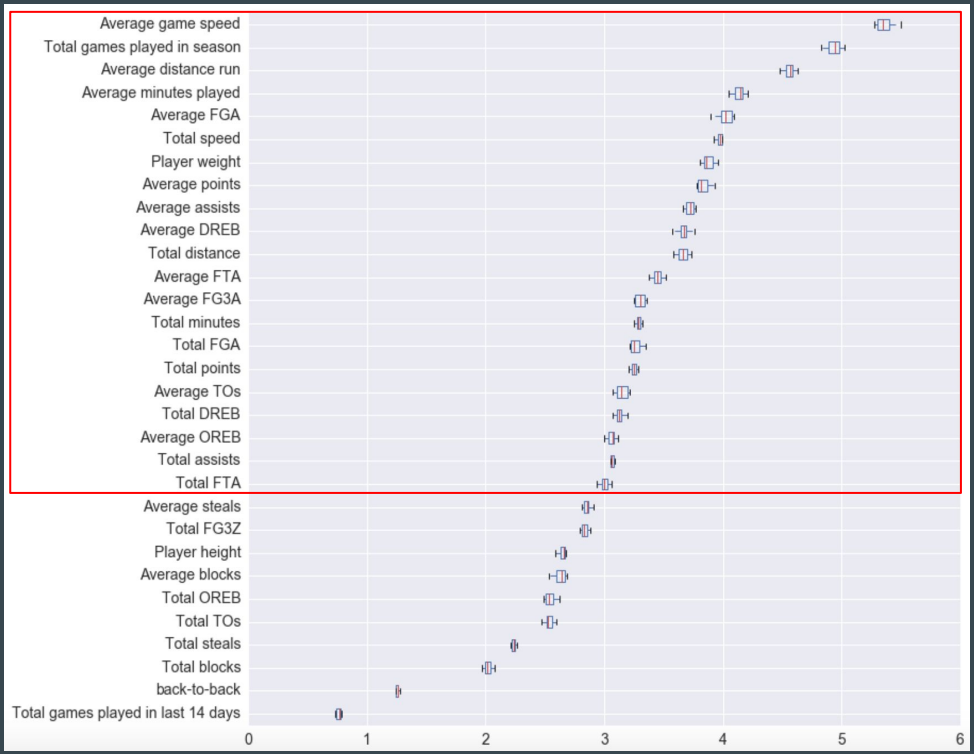
- Sensors in arena
- Important Statistics
  - Avg. Game Speed
  - Avg. Distance Run
  - Post Ups
  - Drives to the Basket
- JSON to CSV converter
- Combine Sheets

SportsVu Data Spreadsheet

PLAYER_ID	PLAYER_NAME	GP	MIN	DIST_FEET	AVG_SPEED	POST_TOUCHES	DRIVES
203932	Aaron Gordon	73	33.78	13504.29	4.28	0	6.2
201143	Al Horford	66	29.09	10815.67	3.98	1.9	1.6
202329	Al-Farouq Aminu	77	28.52	11402.91	4.25	4.6	2.7
202692	Alec Burks	64	21.48	8500.19	4.22	0.2	5.2
203518	Alex Abrines	31	18.98	7906.84	4.38	0.3	1.2
1627936	Alex Caruso	20	17.72	6850.5	4.26	0	3.1
203458	Alex Len	74	19.69	7869.19	4.24	0	0.4
203459	Allen Crabbe	43	26.35	10575.95	4.24	0	2.1
1629019	Allonzo Trier	63	22.78	9013.24	4.17	0.1	7
203083	Andre Drummond	75	33.49	12306.51	3.94	1.3	1.2
2738	Andre Iguodala	66	23.12	8957.11	4.1	0	1
203952	Andrew Wiggins	68	34.83	13285.04	4.04	0.1	8
1627790	Ante Zizic	55	18.11	6814.96	4.01	3	0.3
203076	Anthony Davis	56	33.03	12060.32	3.9	1.4	3.8
201229	Anthony Tolliver	60	16.23	6662.9	4.32	5.9	0.6
203382	Aron Baynes	49	15.92	5985.35	4.01	5.2	0
203085	Austin Rivers	72	26.79	10213	4.06	0.1	5.6
202340	Avery Bradley	63	30.25	12312.37	4.32	0.1	3.3
1628389	Bam Adebayo	78	23.19	9017.58	4.11	0.8	0.8
1627732	Ben Simmons	76	34.43	13337.54	4.17	2.4	9.7
201933	Blake Griffin	72	35.22	12209.85	3.72	0	8.2

- Remove players with less than 15 min a night
- Averages for season up to that point

Scraping Basketball Reference Website



Talukder, H., Vincent, T.: Preventing in-game injuries for NBA players (2016)

- Player Data => Excel Spreadsheet
- link = "https://www.basketball-reference.com/players/"+ (first_letter_of_first_name) + "/" + (first_five_letters_of_last_name) + (first_two_letters_of_first_name) + (version) + "/gamelog/" + year + "/"
- Debugging for players with the same name or nicknames

Adding Injury Data

- Scraped Pro Sports Transactions to put all injury data into a spreadsheet
- Flu, illness, sports hernia, skin infection and other non sports related injuries were removed
- Created a dictionary data structure to store all player injuries

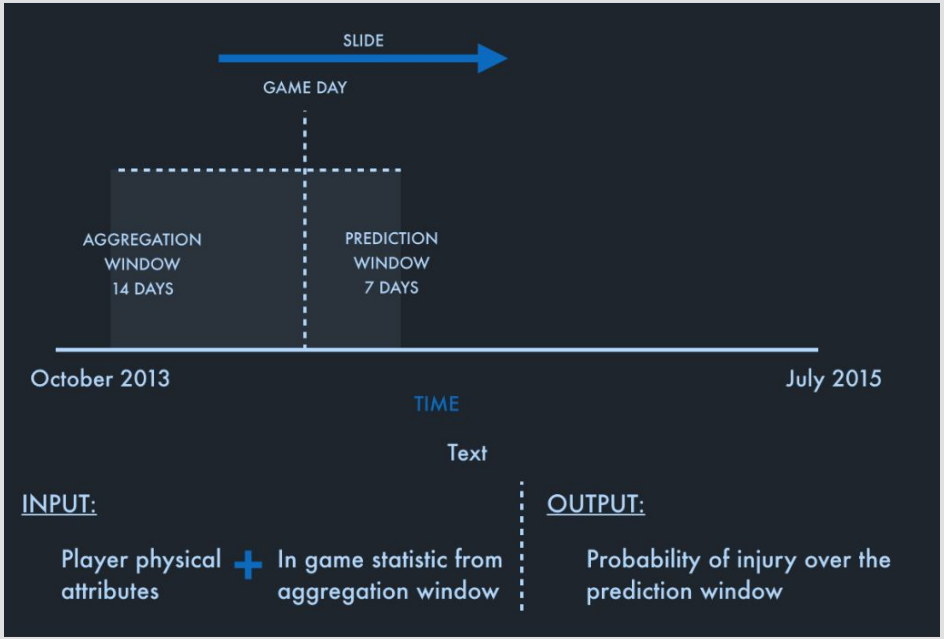
‘Nate Robinson’: {
'2014-02-05': ' recovering from surgery on left knee to repair torn ACL (DNP)',
'2014-02-03': ' recovering from surgery on left knee to repair torn ACL (DNP)',
'2014-01-31': ' torn ACL in left knee (out indefinitely)' }

Player Spreadsheet with Injury Data

GM	Date	Weight	MP	FGA	3PA	FTA	ORB	DRB	AST	TO	Fouls	PTS	dist_feet	avg_speed	post_ups	drives	Injury
33	2014-01-05	180	19.85	11	5	5	3	2	5	0	2	21	7482.474	4.29	0	4.848	
34	2014-01-07	180	22.7	7	1	1	0	0	5	3	2	10	8519.649	4.29	0	5.52	
35	2014-01-09	180	23.13	13	6	1	1	0	2	0	2	6	8741.9	4.29	0	5.664	
36	2014-01-11	180	19.6	4	2	0	0	1	6	0	4	5	7408.39	4.29	0	4.8	
37	2014-01-13	180	16.92	7	3	2	1	3	0	2	2	3	6371.215	4.29	0	4.128	
38	2014-01-15	180	22.85	12	5	3	2	2	3	4	4	24	8593.732	4.29	0	5.568	
39	2014-01-17	180	20.5	11	5	0	0	1	2	0	2	19	7704.726	4.29	0	4.992	
40	2014-01-19	180	24.15	8	4	2	0	1	0	2	2	9	9112.32	4.29	0	5.904	
41	2014-01-23	180	19.85	12	5	2	0	0	1	1	3	13	7482.474	4.29	0	4.848	
42	2014-01-25	180	19	11	5	1	0	0	3	4	2	15	7186.138	4.29	0	4.656	
43	2014-01-26	180	24.38	8	4	4	0	1	8	1	4	12	9186.404	4.29	0	5.952	
44	2014-01-29	180	6.27	2	1	0	0	0	0	0	1	0	2370.685	4.29	0	1.536	torn ACL in left knee (out indefinitely)
	2014-01-31	180	0	0	0	0	0	0	0	0	0	0	0	0	0	0	torn ACL in left knee (out indefinitely)

‘Nate Robinson’: {
'2014-01-31': ' torn ACL in left knee (out indefinitely)' }

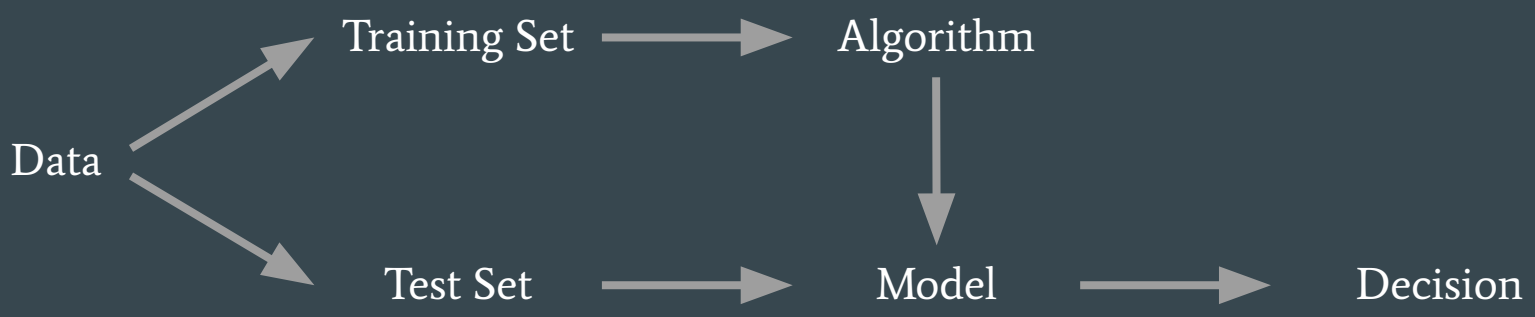
Research Methods



Talukder, H., Vincent, T.: Preventing in-game injuries for NBA players (2016)

- 8 game aggregation window
- 4 game prediction window
- Predictive Performance (longer windows) vs. Actionability (shorter windows)
- Remove windows players missed games for non-injury reasons

Machine Learning



Algorithms Tested

Naive Bayes

- Each feature is given the same weight and no pair of features are dependent

Decision Tree

- Starts with a single feature, then branches off depending on whether that player met that criteria or not, eventually getting to a final result

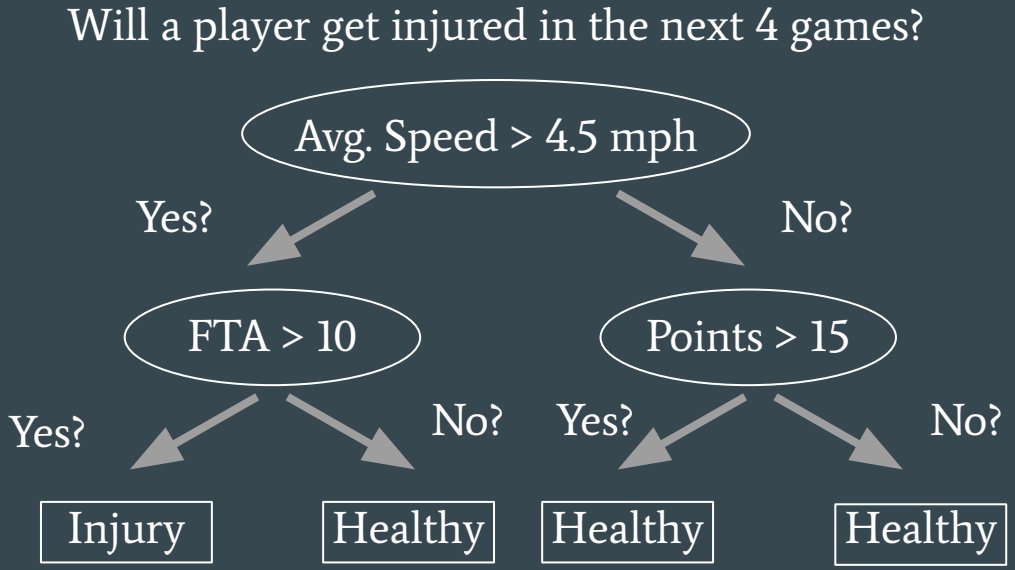
Logistic Regression

- Analyzes a dataset in which there are one or more independent variables that determine a binary outcome

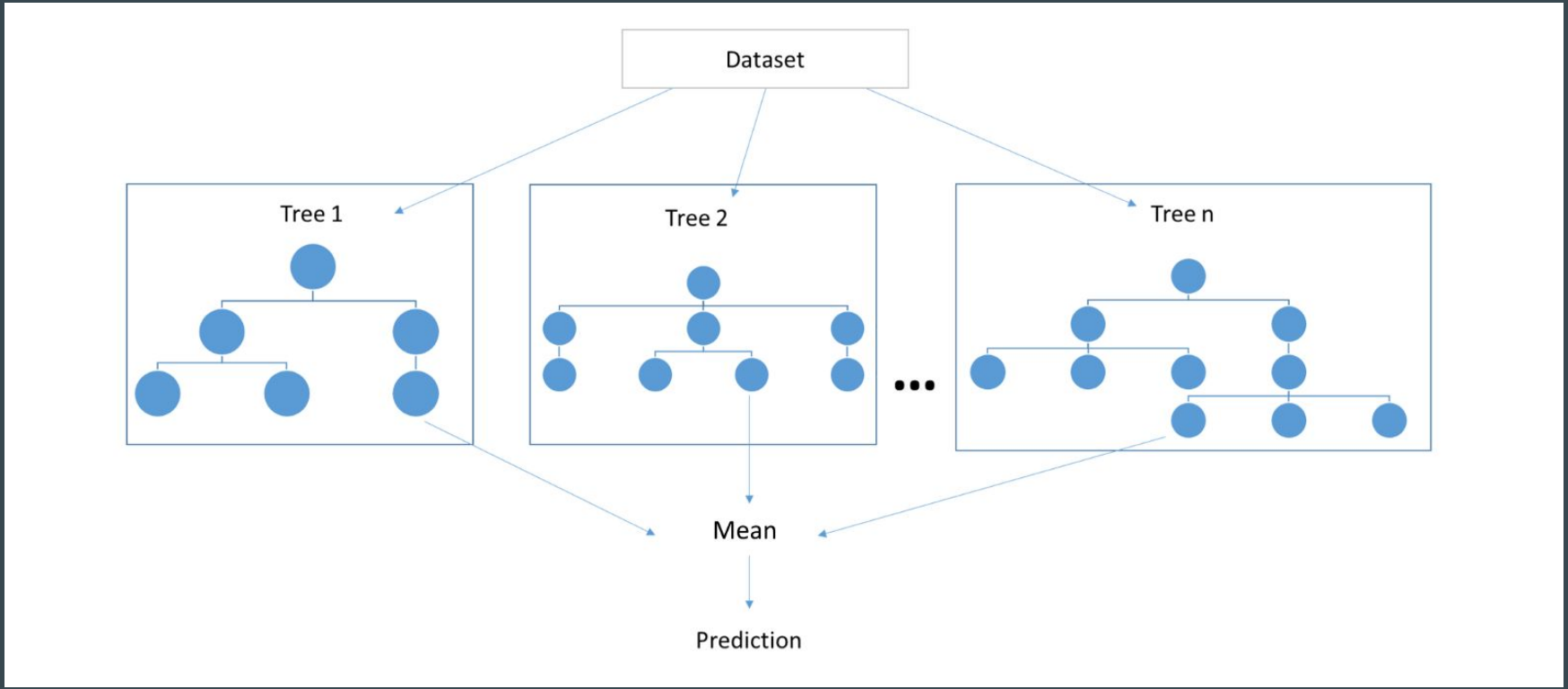
Random Forest Regression Classifier

- Fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting

Decision Tree



Random Forest

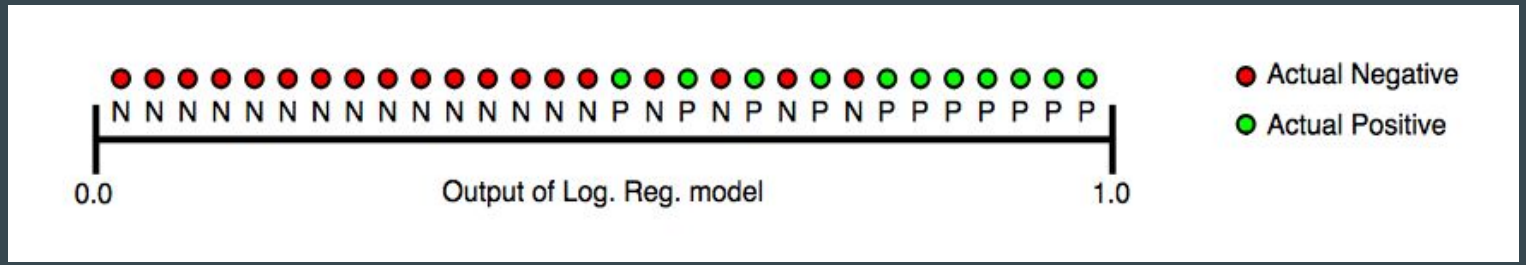


Measuring Results

Accuracy

	precision	recall	f1-score	support
F	0.78	0.97	0.87	342
T	0.96	0.75	0.84	364
avg/total	0.88	0.86	0.85	706

Area Under Curve (AUC)



Results

Algorithm	Accuracy	AUC
Decision Tree	.79	.79
Naive Bayes	.80	.81
Logistic Regression	.82	.87
Random Forest	.86	.90

Nate Robinson 2014 Season Predictions

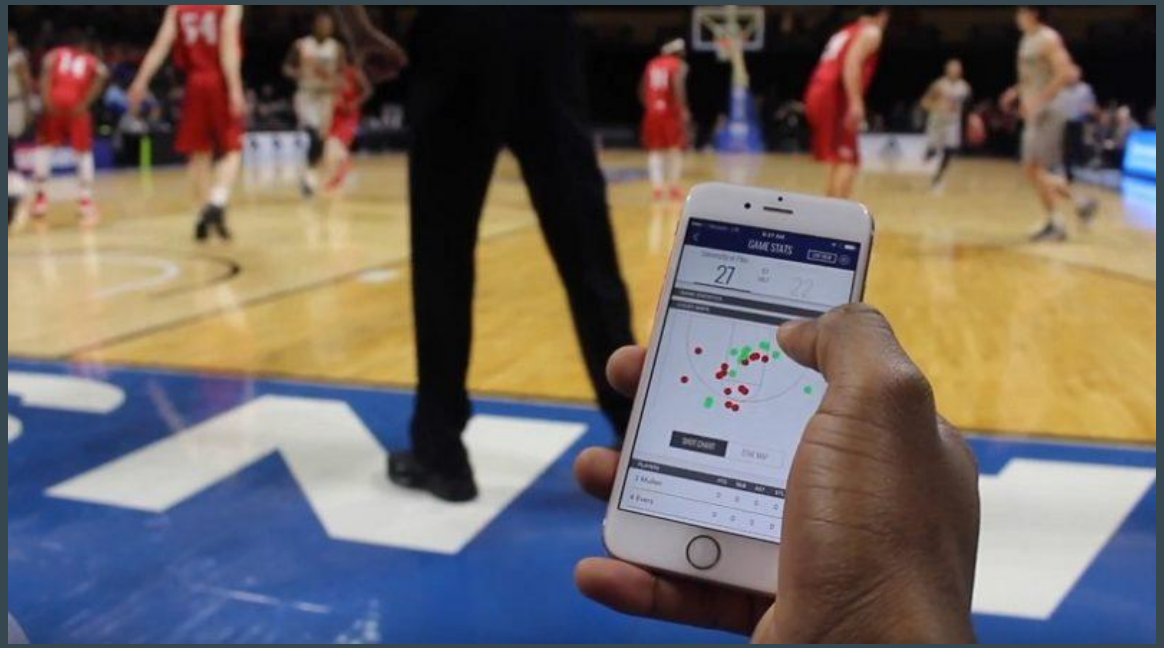
F	F	F	F	F	F
F	F	F	F	F	F
F	F	F	F	F	F
F	F	F	F	F	F
F	F	F	F	F	F
F	F	F	F	F	F
T	T	T	T	T	T
T	T	T	T	T	T
T	T	T	T	T	T

Table 1: Nate Robinson Injury Array

0.33	0.28	0.36	0.09	0.11	0.17
0.15	0.11	0.13	0.26	0.14	0.13
0.16	0.12	0.17	0.12	0.13	0.30
0.18	0.17	0.16	0.12	0.32	0.22
0.11	0.13	0.16	0.25	0.21	0.19
0.07	0.27	0.19	0.20	0.20	0.18,
0.72	0.99	1.00	1.00	1.00	1.00
1.00	1.00	1.00	1.00	1.00	1.00
1.00	1.00	1.00	1.00	1.00	1.00

Table 2: Nate Robinson Predicted Probabilities

Real Time Model



- Cron jobs run every night to update spreadsheets
- Predictions for current season
- On April 5th my model predicted Luka Doncic would get injured at a predicted probability of 0.67
- Injury occurred on April 10th

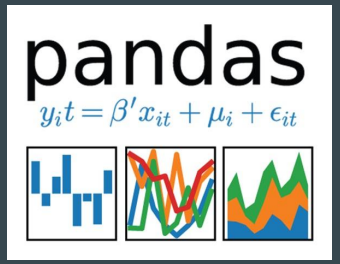
Future Work

- Biometric data
- 24/7 monitoring
- Neural Network
- Expand to other sports

Summary

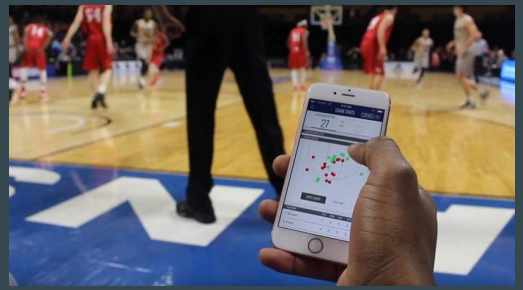
Data Aggregation

Data Organization



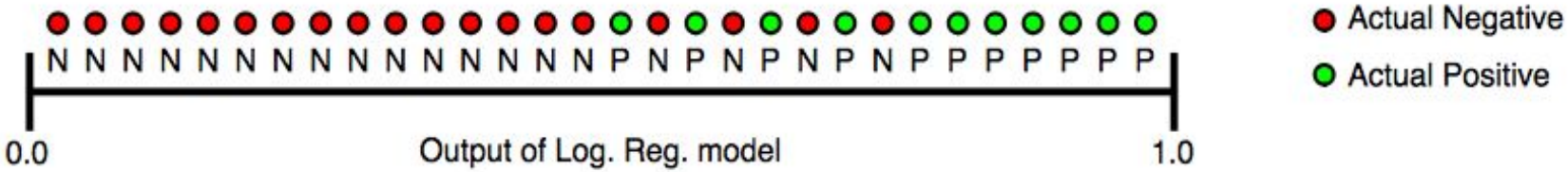
Machine Learning

Real Time Model



Questions?

Area Under ROC Curve



- Determines how good the model is for distinguishing the given classes (injured or not injured), in terms of predicted probability
- 0.0 as the worst to 1.0 as the best
- The receiver operating characteristic curve (ROC), plots both true positive and false positive rate
- One of the most commonly used metrics to evaluate performance of machine learning algorithms

Sliding Window Approach

	A	B	C	D	E	F	G	H	I
1		GM	Date	FGA	FTA	MIN	AVG_DIST_FEET	AVG_SPEED	INJURED
2	0	1	2013-10-30	24	24	39	11434.56	3.85	F
3	1	2	2013-11-01	11	5	27	11434.56	3.85	F
4	2	3	2013-11-03	19	13	39	11434.56	3.85	F
5	3	4	2013-11-06	16	9	36	11434.56	3.85	F
6	4	5	2013-11-08	15	19	44	11434.56	3.85	F
7	5	6	2013-11-10	23	9	46	11434.56	3.85	F
8	6	7	2013-11-13	18	17	41	11434.56	3.85	F
9	7	8	2013-11-14	13	9	42	11434.56	3.85	F
10	8	9	2013-11-16	15	7	38	11434.56	3.85	F
11	9	10	2013-11-18	21	15	37	11434.56	3.85	F
12	10	11	2013-11-21	12	5	39	11434.56	3.85	F
13	11	12	2013-11-24	9	13	26	11434.56	3.85	F
14	12	13	2013-11-27	23	3	38	11434.56	3.85	F
15	13	14	2013-11-29	22	12	43	11434.56	3.85	F
16	14	15	2013-12-01	21	2	44	11434.56	3.85	F
17	15	16	2013-12-03						T
18	16	17	2013-12-04						T
19	17	18	2013-12-06	17	5	36	11434.56	3.85	F
20	18	19	2013-12-08	23	6	34	11434.56	3.85	F
21	19	20	2013-12-10	17	5	36	11434.56	3.85	F
22	20	21	2013-12-11	23	12	38	11434.56	3.85	F

Kevin Durant Data

- 14 game aggregation window
- 7 game prediction window
- Average accuracy of 76%

A game approaches...

What am I doing differently?

	A	B	C	D	E	F	G	H	I
1		GM	Date	FGA	FTA	MIN	AVG_DIST_FEET	AVG_SPEED	INJURED
2	0	1	2013-10-30	24	24	39	11434.56	3.85	F
3	1	2	2013-11-01	11	5	27	11434.56	3.85	F
4	2	3	2013-11-03	19	13	39	11434.56	3.85	F
5	3	4	2013-11-06	16	9	36	11434.56	3.85	F
6	4	5	2013-11-08	15	19	44	11434.56	3.85	F
7	5	6	2013-11-10	23	9	46	11434.56	3.85	F
8	6	7	2013-11-13	18	17	41	11434.56	3.85	F
9	7	8	2013-11-14	13	9	42	11434.56	3.85	F
10	8	9	2013-11-16	15	7	38	11434.56	3.85	F
11	9	10	2013-11-18	21	15	37	11434.56	3.85	F
12	10	11	2013-11-21	12	5	39	11434.56	3.85	F
13	11	12	2013-11-24	9	13	26	11434.56	3.85	F
14	12	13	2013-11-27	23	3	38	11434.56	3.85	F
15	13	14	2013-11-29	22	12	43	11434.56	3.85	F
16	14	15	2013-12-01	21	2	44	11434.56	3.85	F
17	15	16	2013-12-03						Minor
18	16	17	2013-12-04						Minor
19	17	18	2013-12-06	17	5	36	11434.56	3.85	F
20	18	19	2013-12-08	23	6	34	11434.56	3.85	F
21	19	20	2013-12-10	17	5	36	11434.56	3.85	F
22	20	21	2013-12-11	23	12	38	11434.56	3.85	F

- predict severity of injury (minor vs major)
- minor = strains and sprains (out a few games)
- major = tears and breaks (out multiple games or rest of season)

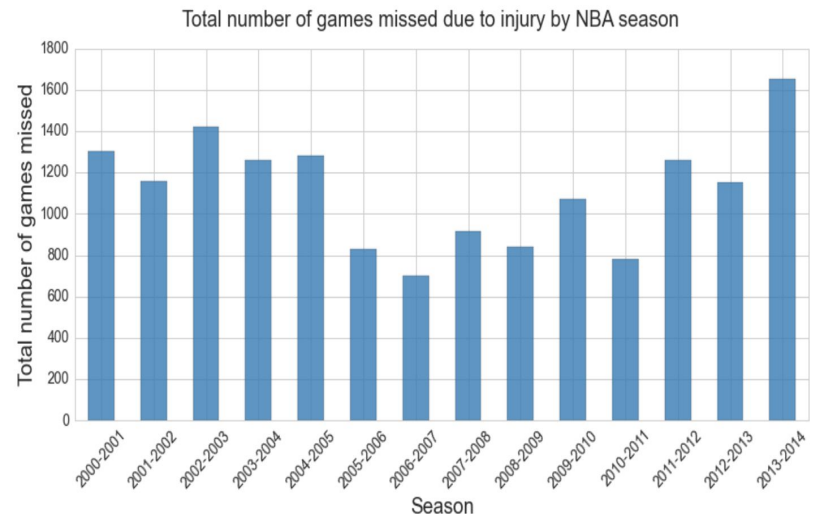
Learning

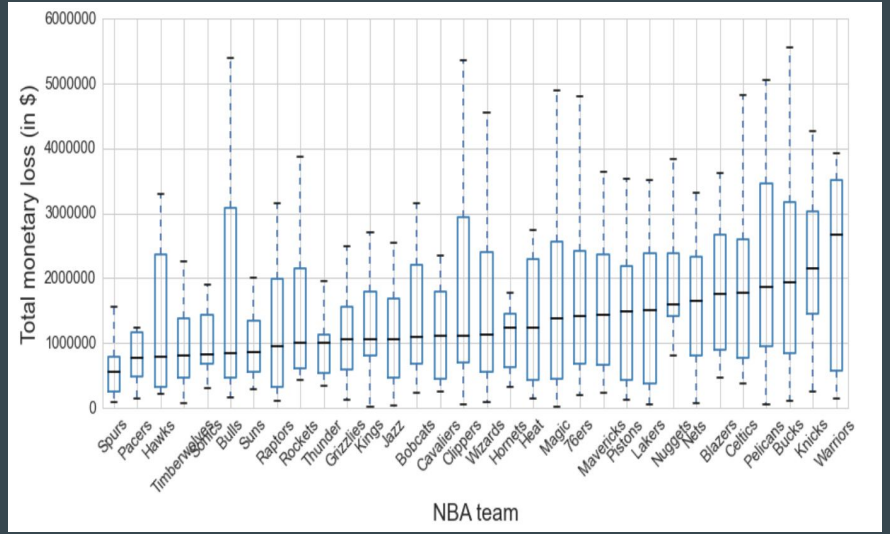
- Web Scraping
  - JSON to CSV
  - Beautiful Soup
- Machine learning
  - Scikit-learn
  - Pandas
  - Algorithms

Improved Windows

22	23	2013-12-15	180	26.58	10	5	2	0	2	4	1	5	14	10001.33	4.29	0	6.48	
23	24	2013-12-17	180	24.68	9	6	2	0	1	2	4	1	12	9260.488	4.29	0	6	
24	25	2013-12-20	180	15.33	4	2	2	1	1	3	0	1	6	5778.544	4.29	0	3.744	
25	26	2013-12-21	180	21.43	7	6	3	1	1	3	2	3	11	8075.145	4.29	0	5.232	
26	27	2013-12-23	180	16.72	9	4	4	1	1	1	0	2	13	6297.132	4.29	0	4.08	
27	28	2013-12-27	180	15.73	3	2	0	0	0	1	3	1	0	5926.712	4.29	0	3.84	
28	29	2013-12-28	180	17.55	8	5	0	1	1	0	4	3	11	6593.467	4.29	0	4.272	
29	30	2013-12-30	180	10.95	1	1	0	0	1	0	2	0	0	4148.698	4.29	0	2.688	
30	31	2014-01-01	180	13.13	3	1	0	0	2	0	0	0	0	4963.621	4.29	0	3.216	
31	32	2014-01-03	180	22.72	10	5	4	1	1	7	1	2	15	8519.649	4.29	0	5.52	
32	33	2014-01-05	180	19.85	11	5	5	3	2	5	0	2	21	7482.474	4.29	0	4.848	
33	34	2014-01-07	180	22.7	7	1	1	0	0	5	3	2	10	8519.649	4.29	0	5.52	
34	35	2014-01-09	180	23.13	13	6	1	1	0	2	0	2	6	8741.9	4.29	0	5.664	
35	36	2014-01-11	180	19.6	4	2	0	0	1	6	0	4	5	7408.39	4.29	0	4.8	
36	37	2014-01-13	180	16.92	7	3	2	1	3	0	2	2	3	6371.215	4.29	0	4.128	
37	38	2014-01-15	180	22.85	12	5	3	2	2	3	4	4	24	8593.732	4.29	0	5.568	
38	39	2014-01-17	180	20.5	11	5	0	0	1	2	0	2	19	7704.726	4.29	0	4.992	
39	40	2014-01-19	180	24.15	8	4	2	0	1	0	2	2	9	9112.32	4.29	0	5.904	
40	41	2014-01-23	180	19.85	12	5	2	0	0	1	1	3	13	7482.474	4.29	0	4.848	
41	42	2014-01-25	180	19	11	5	1	0	0	3	4	2	15	7186.138	4.29	0	4.656	
42	43	2014-01-26	180	24.38	8	4	4	0	1	8	1	4	12	9186.404	4.29	0	5.952	
43	44	2014-01-29	180	6.27	2	1	0	0	0	0	0	1	0	2370.685	4.29	0	1.536	torn ACL in left knee (out indefinitely)
44		2014-01-31	180	0	0	0	0	0	0	0	0	0	0	0	0	0	0	torn ACL in left knee (out indefinitely)

Why are injuries important?





Individual features that contribute to injury events

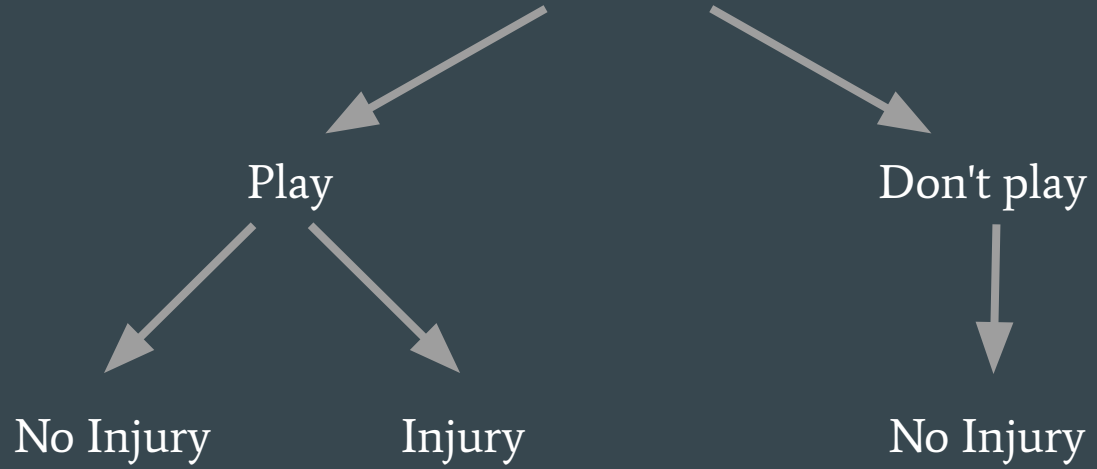
Most Important Features

1. the average speed at which a player ran during games
2. the total number of games played
3. the average distance covered by a player
4. the average number of minutes played
5. the average number of field goals attempted

Non-factor

- number of back to back games

A game approaches...



Data 2014-2015

	A	B	C	D	E	F	G	H	I	J
1	PLAYER_ID	PLAYER_NAME	GP	MIN	DIST_FEET	AVG_SPEED	FGA	DRIVES	FTA	INJURED
2	201977	Marcus Thornton	48	15.01	5443.73	4.14	7.2	2.1	1.2	F
3	202397	Ish Smith	55	15.12	6032.75	4.54	6.7	9.9	1	F
4	202714	Shelvin Mack	55	15.14	6016.18	4.52	5.3	4.6	0.6	F
5	202331	Paul George	25	15.15	5475	4.1	8.2	1.7	1.8	F
6	203956	Mitch McGary	32	15.15	6023.75	4.53	5.2	0.6	1.3	T
7	203798	PJ Hairston	45	15.27	5783.87	4.32	6	1.5	0.8	F
8	203960	JaKarr Sampson	74	15.29	6117.81	4.56	4.7	4.1	1.3	F
9	203088	Kendall Marshall	27	15.33	5678.44	4.21	3.6	4.1	0.3	T
10	203933	T.J. Warren	40	15.35	5942.98	4.4	5.4	1.8	0.5	T
11	203086	Meyers Leonard	55	15.39	5427.44	4.01	4.5	0.4	0.6	F
12	204014	Damjan Rudez	68	15.4	5711.44	4.22	3.9	0.8	0.3	T
13	203917	Nik Stauskas	73	15.43	5871.96	4.33	4.1	2	0.9	T
14	1889	Andre Miller	81	15.47	5387.01	3.96	3.6	3.5	1	F
15	202338	Kevin Seraphin	79	15.63	5739.11	4.18	5.7	0.2	1	T
16	203561	Brandon Davies	27	15.66	6299.85	4.57	5.1	1.2	1.4	F
17	203468	CJ McCollum	62	15.7	6438.69	4.68	5.9	3.6	1.2	F
18	201988	Patty Mills	51	15.71	6653.55	4.83	6.6	2	0.8	F
19	203461	Anthony Bennett	56	15.8	5841.05	4.21	5.2	0.6	1.1	T
20	201578	Marreese Speights	76	15.88	5469.83	3.86	8.5	0.8	2.3	F
21	202332	Cole Aldrich	60	16.03	5715.17	4.05	4.9	0.1	1	T
22	203382	Aron Baynes	70	16.03	5906.49	4.2	4.7	0.4	1.5	T
23	2617	Udonis Haslem	61	16.1	5525.3	3.91	3.9	0.3	1	T
24	203124	Kyle O'Quinn	51	16.15	5627.67	3.97	4.7	0.6	1.1	F
25	2501	Reggie Evans	47	16.25	5774.55	4.04	2.9	0.2	2.1	T
26	203101	Miles Plumlee	73	16.36	6056.18	4.21	3.5	0.2	0.6	F
27	201973	Jonas Jerebko	75	16.4	6301.37	4.38	5.1	1.7	0.9	F
28	203544	Pero Antic	63	16.46	5827.63	4.03	4.7	0.4	2	T
29	203118	Mike Scott	68	16.51	6247.84	4.32	6.7	0.7	1.1	F
30	1713	Vince Carter	66	16.53	5842.29	4.03	6.2	1.9	0.9	F
31	203142	Chris Copeland	50	16.55	5868.56	4.05	6.2	1.7	0.9	F
32	203921	Cleanthony Early	39	16.59	6286.64	4.28	5.5	1.7	1.2	T
33	203900	Markel Brown	47	16.63	6220.34	4.26	4.5	1.7	1.2	F
34	201585	Kosta Koufos	81	16.65	6141.95	4.2	4.4	0.2	1	F

- Predicting whether a player gets injured in a full season
- Remove players with less than 14 games
- Remove players with less than 15 min per game
- Average accuracy of 80%

Derrick Rose: what could have been

- 2008-2009 Rookie of the year
- 2009-2010, 2010-2011, 2011-2012 All-Star
- 2010-2011 MVP
- 2011-2012 Missed Season: ACL injury