Predicting NBA Player Performance

Team members

Patrick Miller, Vandana Ramakrishnan, Nathaniel Matare, Ernie Mori, Jordan Bell-Masterson

Data Set

We have aggregated 4259 features for a dataset that includes 97,866 observations of individual players appearing in the 2013 - 2016 NBA seasons. The dataset is time series dependent; each observation at time t-0 includes lagged statistics leading up to time t-0, and known information at time t-0.

- Player biographical information known at time t-0
- Game atmospherics at time t-0
- Individual player statistics at time t -1 through time t-10
- The player's current injury or lack thereof at time t-0 through time t-10
- The players team statistics at time t -1 through time t-10
- The opposing teams statistics at time t -1 through time t-10
- The 'type' of player [similarity] as determined by season averages via k-means clustering

Additionally, the dataset includes matchup information encoded as [-1, 0, 1]:

- [1] if the player or team is either the player or on the players team
- [0] if the player, team or official was not present during the game
- [-1] if the player or team was opposing the player

Further, for each player we have aggregated news and analyst reports between games played:

- The probability of the text falling into one of twenty topics; as determined by an LDA model.
- N-gram tokenization of forty specific key words and phrases; reported as frequency.

Project Idea:

We will use this dataset to predict the number of points a player will score at time t-0 given information at time t-0 and historical information available until time t-0.

Given the large number of features relative to an individual player observation—at the most ~260 player observations to 4259 features—we will likely employ dimensionality reduction via PCA, MCA, or clustering. Further, because of the time series dependent nature of the dataset, cross validation will be tricky.[1] Thus, we have written several utility functions to perform time-forward cross validation.[2] We hypothesize that we will face a highly interacted environment affected by player idiosyncrasies and game matchups; thus we intend to employ Deep Neural Networks. However, we may find better success with a combination of models and are therefore considering ensemble models: a combination of trees and regularized regressions. Additionally, we may have several sub-prediction problems; that is, predicting win probability may help us better estimate a players point prediction.

Please see the appendix for additional information.

[1] http://robjhyndman.com/hyndsight/tscv/

[2] https://github.com/nmatare

Appendix:

Key								
.stat_player [- ∞, ∞]	logical	game_starter	.stat_player_team [- ∞, ∞]	numeric	team_points	.stat_opposing_team [- ∞, ∞]	numeric	team_points
	logical	game_played		numeric	field_goals_pct		numeric	field_goals_pct
	numeric	field_goals_pct		numeric	three_points_pct		numeric	three_points_pct
	numeric	three_points_pct		numeric	two_points_pct		numeric	two_points_pct
	numeric	two_points_pct		numeric	free_throws_pct		numeric	free_throws_pct
	numeric	free_throws_pct		integer	rebounds		integer	rebounds
	integer	rebounds		integer	assists		integer	assists
	integer	assists		integer	turnovers		integer	turnovers
	integer	turnovers		integer	steals		integer	steals
	integer	steals		integer	blocks		integer	blocks
	integer	blocks		numeric	assists_turnover_ratio		numeric	assists_turnover_ratio
	numeric	assists_turnover_ratio		numeric	personal_fouls		numeric	personal_fouls
	integer	personal_fouls		numeric	team_tech_fouls		numeric	team_tech_fouls
	integer	tech_fouls		numeric	flagrant_fouls		numeric	flagrant_fouls
	integer	flagrant_fouls		numeric	pls_mins		numeric	pls_mins
	numeric	pls_mins		integer	points		integer	points
	integer	points		numeric	minutes		numeric	minutes
	numeric	minutes		numeric	fast_break_pts		numeric	fast_break_pts
		points_leader		numeric	paint_pts		numeric	paint_pts
		rebounds_leader		numeric	team_turnovers		numeric	team_turnovers
		assists_leader		numeric	team_rebounds		numeric	team_rebounds
	logical	win						
.bio [0, ∞]	numeric	height	.text [0, ∞]	numeric	report	.injury [0,1]	categorical	Hand
	numeric	weight		numeric	swarm			Illness
	numeric	experience	.text_ngram [0, ∞]	integer	hot streak			Quad
	categorical	college		integer	extra minutes			Heart
	categorical	position		integer	momentum			…
	categorical	primary_position		integer	…	.desc [0,1]	categorical	game_time
	categorical	game_position	.matchup [-1, 0, 1]	integer	Lebron James		categorical	game_month
	categorical	jersey_number		integer	Anthony Davis		categorical	game_day
	categorical	status		integer	Josh Smith		categorical	game_year
				integer	Jose Calderon		categorical	game_title
				integer	…		categorical	venue_name
			.team_matchup[-1, 0, 1]	integer	GoldenStateWarriors		categorical	broadcast_network
				integer	BostonCeltics.team_matchup		logical	home
				integer	BrooklynNets.team_matchup	.player_season_average [0,1]	integer	cluster_1
				integer	MinnesotaTimberwolves.team_matchup		integer	cluster_2
				integer	WashingtonWizards.team_matchup		integer	cluster_3
				integer	…		integer	cluster_4
			.officials[-1, 0, 1]	integer	Donald Hudson_as_Official		integer	cluster_5
				integer	Ken Mauer_as_Head Official		integer	cluster_6
				integer	Curtis Blair_as_Alternate.official		integer	cluster_7
				integer	Mark Lindsay_as_Head Official.official		integer	cluster_8
				integer	…		integer	…