PREDICTING NBA REGULAR SEASON WINS USING PYTHON

Shaun Chaudhary (2016)

OVERVIEW

- Build a model to predict the upcoming regular season win percentage of an NBA team given their previous season's statistics

- Many white papers have been written to try and develop a robust predictive model. Some techniques used:
  - Linear Regression – *fivethirtyeight.com* and *Giarta, E et al. (2015)*
  - Naïve Bayes Classification – *Na Wei (2011)*
  - Deep Learning / Neural Networks – *Bernard, L et al. (2009)*

- NBA Base Knowledge
  - 82 games in a regular season
  - 48 minutes in a game (12 minutes / quarter)
  - 5 players on the court per team

THE PROCESS

1. Identify data sources that have historical, clean NBA data
    1. www.basketball-reference.com
    2. Scraped data from the first few tables using pandas (*pd.read_html*) and manually copied and pasted the rest of the data into csv files that are stored on my local hard drive

2. Read into python and clean up data types / column headers
    1. Convert numeric columns, remove miscellaneous unicode characters, and standardize column headers across each table

3. Merge individual tables of data on team name and year
    1. Combine historical standings, offensive, defensive, all star, and first team all-nba data into a single DataFrame

4. Log transform features with continuous data and build Kernel Ridge Regression with 10-fold validation

RAW DATA EXAMPLES

Conference Standings

* Playoff teams

Eastern Conference	W	L	W/L%	GB	PS/G	PA/G	SRS
Cleveland Cavaliers* (1)	57	25	.695	—	104.3	98.3	5.45
Toronto Raptors* (2)	56	26	.683	1.0	102.7	98.2	4.08
Miami Heat* (3)	48	34	.585	9.0	100.0	98.4	1.50
Atlanta Hawks* (4)	48	34	.585	9.0	102.8	99.2	3.49
Boston Celtics* (5)	48	34	.585	9.0	105.7	102.5	2.84
Charlotte Hornets* (6)	48	34	.585	9.0	103.4	100.7	2.36
Indiana Pacers* (7)	45	37	.549	12.0	102.2	100.5	1.62
Detroit Pistons* (8)	44	38	.537	13.0	102.0	101.4	0.43
Chicago Bulls (9)	42	40	.512	15.0	101.6	103.1	-1.46
Washington Wizards (10)	41	41	.500	16.0	104.1	104.6	-0.50
Orlando Magic (11)	35	47	.427	22.0	102.1	103.7	-1.68
Milwaukee Bucks (12)	33	49	.402	24.0	99.0	103.2	-3.98
New York Knicks (13)	32	50	.390	25.0	98.4	101.1	-2.74
Brooklyn Nets (14)	21	61	.256	36.0	98.6	106.0	-7.12
Philadelphia 76ers (15)	10	72	.122	47.0	97.4	107.6	-9.92

Western Conference	W	L	W/L%	GB	PS/G	PA/G	SRS
Golden State Warriors* (1)	73	9	.890	—	114.9	104.1	10.38
San Antonio Spurs* (2)	67	15	.817	6.0	103.5	92.9	10.28
Oklahoma City Thunder* (3)	55	27	.671	18.0	110.2	102.9	7.09
Los Angeles Clippers* (4)	53	29	.646	20.0	104.5	100.2	4.13
Portland Trail Blazers* (5)	44	38	.537	29.0	105.1	104.3	0.98
Dallas Mavericks* (6)	42	40	.512	31.0	102.3	102.6	-0.02
Memphis Grizzlies* (7)	42	40	.512	31.0	99.1	101.3	-2.14
Houston Rockets* (8)	41	41	.500	32.0	106.5	106.4	0.34
Utah Jazz (9)	40	42	.488	33.0	97.7	95.9	1.84
Sacramento Kings (10)	33	49	.402	40.0	106.6	109.1	-2.32
Denver Nuggets (10)	33	49	.402	40.0	101.9	105.0	-2.81
New Orleans Pelicans (12)	30	52	.366	43.0	102.7	106.5	-3.56
Minnesota Timberwolves (13)	29	53	.354	44.0	102.4	106.0	-3.38
Phoenix Suns (14)	23	59	.280	50.0	100.9	107.5	-6.32
Los Angeles Lakers (15)	17	65	.207	56.0	97.3	106.9	-8.92

RAW DATA EXAMPLES (CTD.)

Team Stats

* Playoff teams   Share & more ▼   Glossary

Rk	Team	G	MP	FG	FGA	FG%	3P	3PA	3P%	2P	2PA	2P%	FT	FTA	FT%	ORB	DRB	TRB	AST	STL	BLK	TOV	PF	PTS	PS/G
1	Golden State Warriors*	82	19880	3489	7159	.487	1077	2592	.416	2412	4567	.528	1366	1790	.763	816	2972	3788	2373	689	498	1245	1701	9421	114.9
2	Oklahoma City Thunder*	82	19830	3372	7082	.476	678	1945	.349	2694	5137	.524	1616	2067	.782	1071	2916	3987	1883	603	487	1305	1691	9038	110.2
3	Sacramento Kings	82	19805	3283	7083	.464	660	1839	.359	2623	5244	.500	1514	2089	.725	868	2760	3628	2009	733	368	1326	1676	8740	106.6
4	Houston Rockets*	82	19830	3094	6847	.452	878	2533	.347	2216	4314	.514	1671	2407	.694	930	2601	3531	1821	821	430	1307	1790	8737	106.5
5	Boston Celtics*	82	19780	3216	7318	.439	717	2142	.335	2499	5176	.483	1520	1929	.788	950	2733	3683	1981	752	348	1127	1796	8669	105.7
6	Portland Trail Blazers*	82	19805	3167	7040	.450	864	2336	.370	2303	4704	.490	1424	1889	.754	948	2782	3730	1748	562	380	1200	1782	8622	105.1
7	Los Angeles Clippers*	82	19830	3141	6759	.465	797	2190	.364	2344	4569	.513	1490	2152	.692	721	2727	3448	1873	709	460	1063	1746	8569	104.5
8	Cleveland Cavaliers*	82	19855	3171	6888	.460	880	2428	.362	2291	4460	.514	1333	1783	.748	873	2777	3650	1861	551	317	1114	1666	8555	104.3
9	Washington Wizards	82	19755	3238	7033	.460	709	1983	.358	2529	5050	.501	1349	1849	.730	743	2688	3431	2005	708	323	1186	1708	8534	104.1
10	San Antonio Spurs*	82	19705	3289	6797	.484	570	1518	.375	2719	5279	.515	1342	1672	.803	770	2831	3601	2010	677	485	1071	1433	8490	103.5
11	Charlotte Hornets*	82	19855	3036	6922	.439	873	2410	.362	2163	4512	.479	1534	1941	.790	734	2869	3603	1778	595	438	1030	1487	8479	103.4
12	Atlanta Hawks*	82	19830	3168	6923	.458	815	2326	.350	2353	4597	.512	1282	1638	.783	679	2772	3451	2100	747	486	1226	1570	8433	102.8
13	Toronto Raptors*	82	19780	3006	6669	.451	708	1915	.370	2298	4754	.483	1702	2190	.777	836	2724	3560	1536	636	449	1073	1610	8422	102.7
14	New Orleans Pelicans	82	19780	3153	7040	.448	702	1951	.360	2451	5089	.482	1415	1823	.776	782	2712	3494	1818	633	342	1102	1713	8423	102.7
15	Minnesota Timberwolves	82	19880	3095	6668	.464	455	1347	.338	2640	5321	.496	1753	2213	.792	821	2587	3408	1916	656	375	1231	1696	8398	102.4
16	Dallas Mavericks*	82	20005	3064	6900	.444	806	2342	.344	2258	4558	.495	1454	1831	.794	751	2781	3532	1813	560	306	1047	1595	8388	102.3
17	Indiana Pacers*	82	19880	3142	6985	.450	663	1889	.351	2479	5096	.486	1430	1872	.764	847	2779	3626	1741	742	391	1219	1641	8377	102.2
18	Orlando Magic	82	19905	3242	7120	.455	636	1818	.350	2606	5302	.492	1249	1649	.757	843	2709	3552	1933	673	417	1155	1701	8369	102.1
19	Detroit Pistons*	82	19880	3111	7087	.439	740	2148	.345	2371	4939	.480	1399	2095	.668	1021	2777	3798	1594	573	304	1110	1557	8361	102.0
20	Denver Nuggets	82	19830	3093	7003	.442	656	1943	.338	2437	5060	.482	1513	1974	.766	941	2718	3659	1858	609	395	1202	1723	8355	101.9

MODEL

1. Chose to use Polynomial Kernel Ridge Regression with a log transformation
    1. Kernel Ridge Regression combines a normal Ridge Regression and attempts to use the "kernel trick" to reduces the variance of the prediction results.
    2. Ideal because we want the predicted win percentage for each team to be between 0 – 1.

2. Adding degrees to the linear regression (converting to polynomial)
    1. Optimizing for mean squared error (MSE), I attempted to build features that accounted for polynomial transformations of degrees 1, 2, 3, and 4.
    2. Degree == 2 had the best combination of lowest training MSE and test MSE

3. Log transforming the numerical features
    1. Distribution of data is clustered around certain percentages and per game averages
    2. Log transforming the data allowed the data to be more accurately distributed between 0 and 1
        1. For example, FG% of all 32 teams not normally distributed from 0-1 but clustered from 30-45%. Log transforming the data allows us to create a more accurate distribution of the data

DATA DICTIONARY

Feature	Feature Name	Data Type	Description	Log
2-Point FG %	*twoP_Perc*	float	Percent of 2 point field goals made in a season	✔
Opp. 2-Point FG %	*twoP_Perc_opp*	float	Percent of opponent's 2 point field goals made in a season	✔
3-Point FG %	*threeP_Perc*	float	Percent of 3 point field goals made in a season	✔
Opp. 3-Point FG %	*threeP_Perc_opp*	float	Percent of opponent's 3 point field goals made in a season	✔
Free Throw %	*FT_Perc*	float	Percent of free throws made in a season	✔
Playoff Appearance	*playoff_appearance*	binary	1 if qualified for playoffs	
# of All Stars	*all_star_count*	integer	Number of all stars for that team	
# of 1st Team All-NBA	*all_nba_count*	integer	Number of players that get selected for 1st team all NBA	
Off. Rebounds / Game	*ORB_G*	float	Average offensive rebounds per game in a season	✔
Steals / Game	*STL_G*	float	Average steals per game in a season	✔
Turnovers / Game	*TOV_G*	float	Average turnovers per game in a season	✔
Personal Fouls / Game	*PF_G*	float	Average personal fouls committed per game in a season	✔
Opp. Off. Rebounds / Game	*ORB_opp_G*	float	Average offensive rebounds per game by opponents in a season	✔
Opp. Personal Fouls / Game	*PF_opp_G*	float	Average opponent personal fouls committed per game in a season	✔
Next Year Win Percentage	*next_year_wl_perc*	float	Target value: next season's regular season win percentage	

OPTIMIZING THE MODEL

1. Tried a Random Forest Regression with *n_estimators* = 50 and *k-folds* = 3. Results were a little better than 50/50 chance.

```
Random Forest Results:
CV AUC [ 0.69536482  0.74823411  0.30539638], Average AUC 0.582998436645731
```

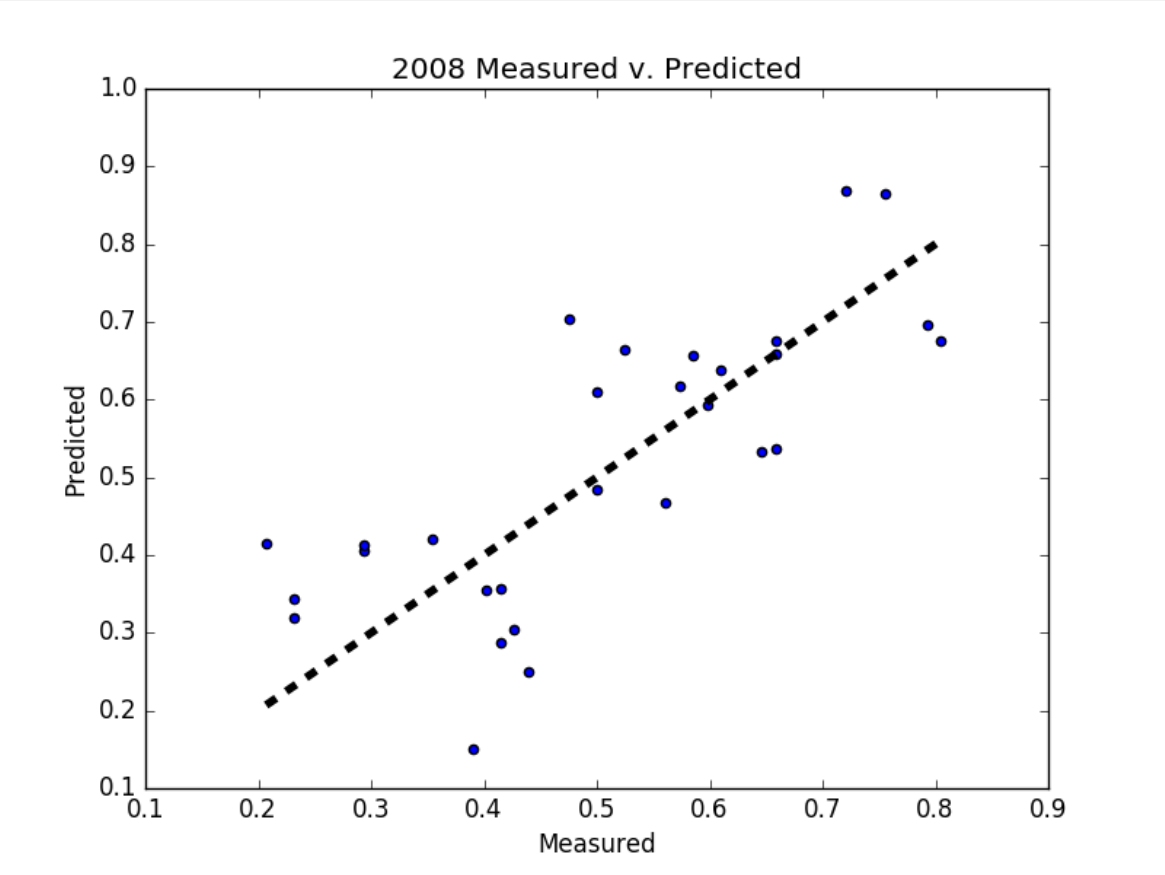
2. Focused on optimizing mean squared error for a polynomial Kernel Ridge Regression with log transformation:

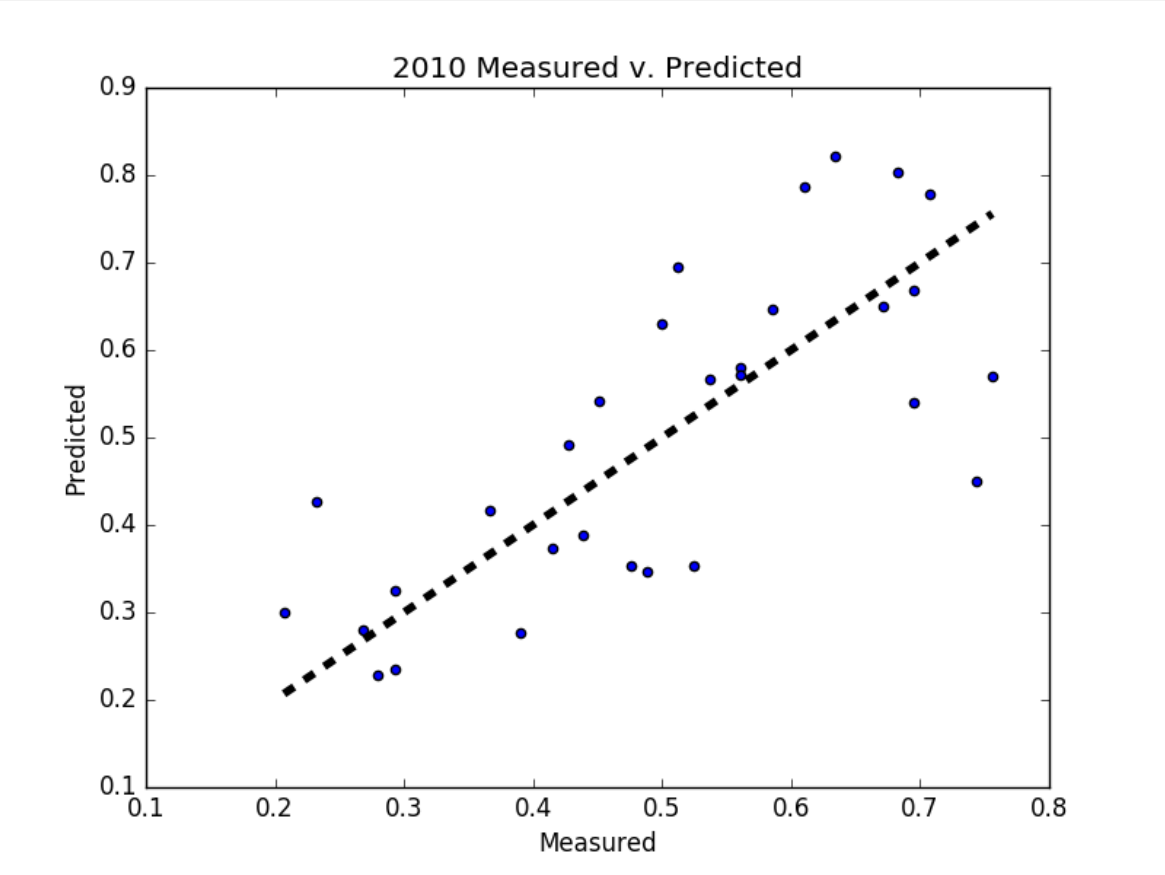
```
Polynomial Regression Results:
Train MSE = 9.40e-03(+/- 5.03e-03)
```

```
2005: Test MSE = 1.82e-01(+/- 1.04e-01)
2006: Test MSE = 2.71e-02(+/- 7.32e-03)
2007: Test MSE = 7.06e-02(+/- 2.84e-02)
2008: Test MSE = 2.66e-02(+/- 1.35e-02)
2009: Test MSE = 4.67e-02(+/- 2.75e-02)
2010: Test MSE = 4.22e-02(+/- 3.83e-02)
2011: Test MSE = 3.47e-02(+/- 7.66e-03)
2012: Test MSE = 4.22e-02(+/- 5.62e-03)
2013: Test MSE = 3.46e-02(+/- 8.32e-03)
2014: Test MSE = 2.49e-02(+/- 9.39e-03)
2015: Test MSE = 5.02e-02(+/- 3.86e-02)
```

- Average prediction delta over 10 years is 10.06 games per team per season

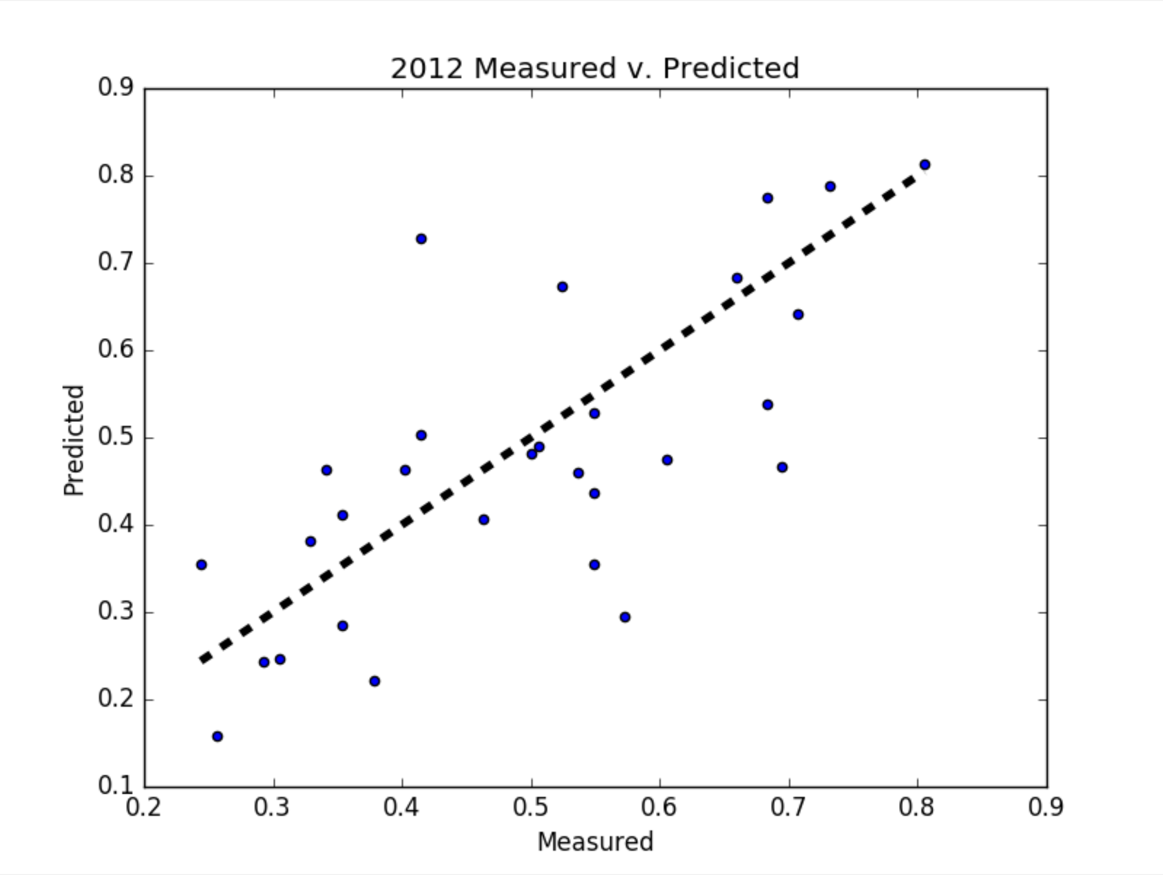
EVALUATING RESULTS

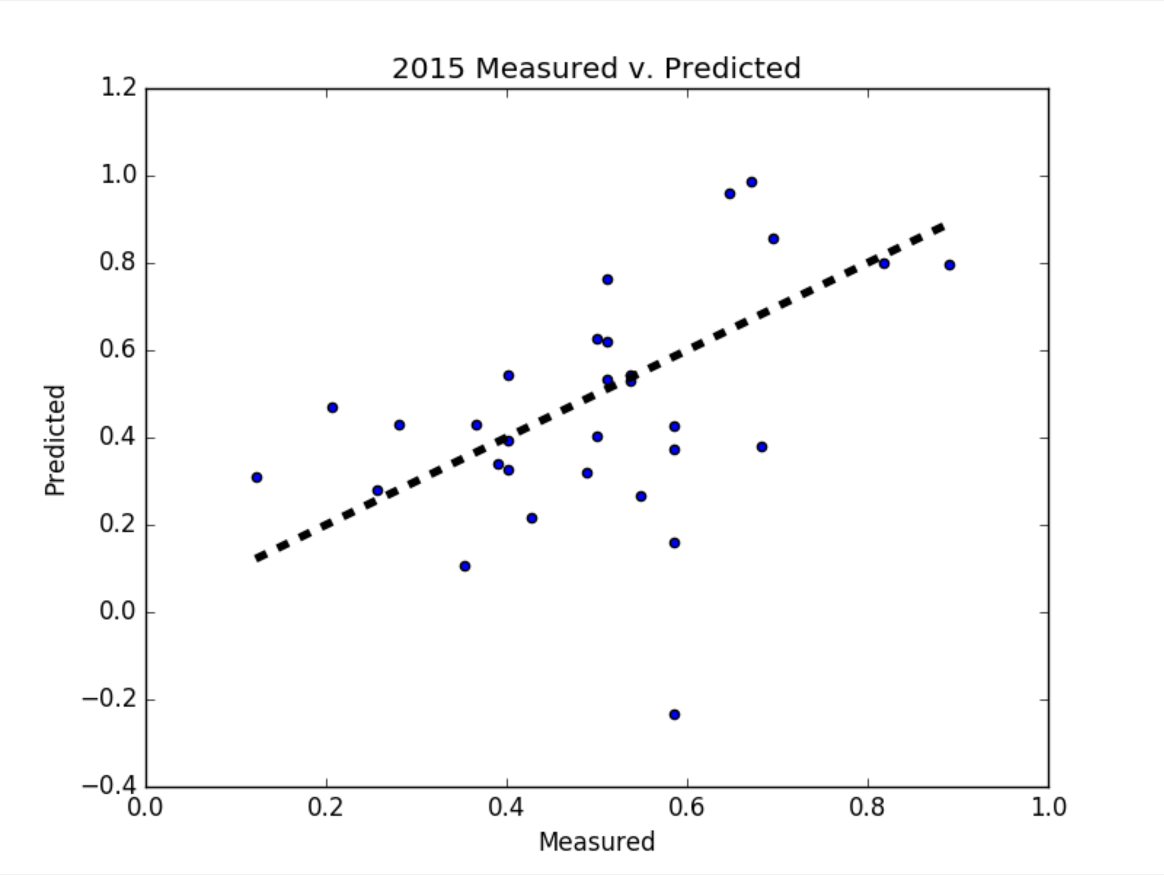




Note: a dot that falls exactly on the dotted line indicates a perfect prediction.

EVALUATING RESULTS





Note: a dot that falls exactly on the dotted line indicates a perfect prediction.

PREDICT 2017 REGULAR SEASON

Eastern Conference	Win %	Games Won
Toronto Raptors	0.618	51
Boston Celtics	0.595	49
Cleveland Cavaliers	0.586	48
Atlanta Hawks	0.580	48
Indiana Pacers	0.578	47
Charlotte Hornets	0.549	45
Miami Heat	0.544	45
Detroit Pistons	0.539	44
Chicago Bulls	0.403	33
Washington Wizards	0.398	33
New York Knicks	0.394	32
Orlando Magic	0.383	31
Milwaukee Bucks	0.376	31
Brooklyn Nets	0.344	28
Philadelphia 76ers	0.321	26

Western Conference	Win %	Games Won
Golden State Warriors	0.677	56
San Antonio Spurs	0.668	55
Oklahoma City Thunder	0.611	50
Los Angeles Clippers	0.601	49
Houston Rockets	0.545	45
Dallas Mavericks	0.529	43
Memphis Grizzlies	0.520	43
Portland Trail Blazers	0.501	41
Utah Jazz	0.412	34
New Orleans Pelicans	0.402	33
Sacramento Kings	0.392	32
Minnesota Timberwolves	0.391	32
Denver Nuggets	0.390	32
Phoenix Suns	0.359	29
Los Angeles Lakers	0.336	28

CONCLUSIONS

- Decent success building a polynomial regression model
  - Using the kernel ridge variety helped limit the predictions to between 0 and 1 (for the most part)
  - Using a log transformation allowed the distribution of normal NBA statistics to be more complete as opposed of clustering around average values

- The average of a 10 game delta for predicting results over 10 years is not terrible given the lack of granularity for each team and the fact we are not accounting for end of season movement of players

- Drawbacks of *sklearn* library in python is that I am unable to evaluate the statistical significance of individual features when using the pipeline functionality

NEXT STEPS / EXTENSIONS

- Need to get more granular with my data and get same statistics but by player
  - This will allow me to better account for the future season's resulting win percentage by accounting for free agency, trades, and retirements
- Become more comfortable evaluating the statistical significance of coefficients when using the *sklearn* library
- Attempt to build logistic regression that will evaluate probability that a team will win an in-season matchup and then aggregate end of season wins
  - Will allow additional analysis into individual matchups between all the teams
- Try implementing a deep learning algorithm / neural network to create a predictive model