William Dreese
Mr. Potter
Final Project Write-Up

Naive Phillies

Baseball has always been referred to as a “numbers game”. What does this mean? It means that each statistic that goes along with the game itself, ranging from the general stats like ERA and Batting Average, to more advanced stats like WAR (wins above replacement) and ballpark numbers (using math and stats to come up with a number 90-110, with a number above 100 deemed a “hitter's” park, and below 100 deemed a “pitcher’s” park), are kept track of and used to rate players and teams. One of my goals of this project was to steer away from these numbers, as they tend to fluctuate over the course of a full season, and to focus on the things that most managers and team owners don’t. My main goal was simple: using a team’s (2015 Phillies) regular season stats for the first 100 games, what position will they end up in after game 162, and what will their record be? I also wanted a system accurate enough to make Pete Rose proud.

The original data set I pulled from baseball-reference.com was looking at the Phillies 2015 season game-by-game. It included winning pitcher, losing pitcher, date, total game time, and many other cool stats. I decided that pitchers didn’t matter over the course of a season, because the idea behind having a rotation is that your ace might face the other team’s 5th man one game, and then your 5th man might also face another team’s ace, leading to a balance over the course of a season. Total game time could be affected by way too many factors that wouldn’t affect the actual score of the game, so that was removed. Other data points that I didn’t include were date, games back, current streak, and score of each game. I decided to not keep the score of each game because I thought it would skew the output. For example, losing to a team by 12 might cause the probability to get too low and then we could technically ‘never’ beat that team. This comes back around to the idea to the pitching rotation and the constant carousel of pitchers that are used throughout the season and the matchups the concept creates.

The five data points I decided to use were Day/Night game, Interleague game, Divisional game, Home/Away, and team played. Day/Night was the first attribute that I focused on. In 2015, the Phils played 53 day games and 109 night games. The Phillies went 23-30 (.434) during the day and 40-69 (.367) under the lights. I thought the .067 difference was important enough to include in the attributes. The next aspect I focused on was Interleague play. I think this attribute of the game is important because either a) a pitcher that rarely gets to bat has to bat OR b) an NL pitcher has to face a DH and not the opposing pitcher, resulting in a score that may not be truly representative of the pitcher’s overall season. Why didn’t I just get rid of these games right away? Because IL games are a part of baseball, and getting rid of 20 games out of a 162 game season (12.3% of games) seemed a bit too much (for reference, the Phils won 8 of 20 games). Up next is one of the most important aspects of a baseball team’s season,

Divisional play. Against Atlanta, Washington, Miami, and New York (Mets) we went 30-44 (.405).

Since defeating your rivals not only gives you a win but also helps you reach the playoffs, this was an obvious choice. It also affects your lineups (ex: giving a good player a rest day during an IL game so he's ready to go against a divisional opponent). Next up is Home/Away. This one definitely stood out as the Phils won 13.6% more games at Citizens Bank Park than away (H: 37-44, .457 A: 26-55, .321). 'Home field advantage' isn't just a saying, and is actually a huge factor when it comes to baseball. Physically, knowing the park you're playing in (fence length, fast/slow grass, etc) as well as having the fans there supporting you makes a huge difference in play. Last but not least, opponent. This doesn't really need much explaining. Different teams have different players that yield different results. As a little factoid to end this intro, the Brewers whooped us 7 games to none over the course of a season, while we outplayed the Cubs (World Series favorites for a while) five games to two, one game of which was Cole Hamels' no-hitter.

I made very basic changes to the code in two places. None of the changes touched the actual algorithm. My first change was for the test/training data. Instead of ⅔ of the data randomly being assigned to training, I set it so that the first 100 games are set to training, and the last 62 are set to testing. I did this easily by setting the variable "trainSize" to 100, as opposed to .67 * the size of the dataset. My second change to the code was where the program processes accuracy. I still keep records on the accuracy the code generates, but the one I was after was win-loss numbers. I simply put an if/then statement where the code is running through all the predictions to judge if the code generated a win or a loss, regardless of the correctness or not.

Class Distribution Charts (Playoff/Missed will be explained later as one of my changes)

	Played	Won	Win %
Home	81	37	0.457
Away	81	26	0.321
Day	53	20	0.434
Night	109	40	0.367
IL	20	8	0.400
NL	143	55	0.385
Out Div	88	33	0.370
In Div	74	30	0.405
Playoff	54	20	0.370
Missed	108	43	0.398

Made By Me

	Opponent				
Split	W	L	RS	RA	WP
ARI	4	2	28	40	.667
ATL	8	11	66	71	.421
BAL	1	3	9	30	.250
BOS	1	5	15	38	.167
CHC	5	2	39	29	.714
CIN	2	4	22	36	.333
COL	2	5	20	31	.286
LAD	2	5	31	39	.286
MIA	10	9	79	75	.526
MIL	0	7	23	41	.000
NYM	5	14	76	111	.263
NYY	2	1	24	24	.667
PIT	2	5	15	21	.286
SDP	5	1	28	21	.833
SFG	1	5	24	43	.167
STL	2	5	28	50	.286
TBR	2	1	10	8	.667
TOR	2	2	17	22	.500
WSN	7	12	72	79	.368

From baseball-reference.com

Statistical Analysis (Mean/ Standard Dev)

Attribute	Mean	Std Dev	Win/Loss	Mean	Std Dev
H/@	1.378	0.492		1.587	0.496
Opponet	7.135	4.9		7.349	4.645
Day/Nigh	1.568	0.502		1.571	0.499
NL/IL	1.162	0.374		1.111	0.317
In/Out	1.405	0.498		1.397	0.493
Playoff?	0.216	0.417		0.349	0.481

The chart above contains the mean and standard deviation for the games won and loss. The integer values I used for each attribute are...

WSN	1	SFG	10
NYM	2	ARI	11
MIA	3	COL	12
ATL	4	BAL	13
--		LAD	14
CHC	5	NYY	15
PIT	6	MIL	16
BOS	7	TBR	17
STL	8	SDP	18
CIN	9	TOR	19

Attribute	Values Used In Data
Home/Away	1,2
Day/Night	1,2
NL/IL	1,2
In Div/ Out of Div	1,2
Playoff/Missed	1,0

The first time I ran my dataset (after turning all values to integers), I got a last 62-game record of 60-2. I knew right away that this was wrong. I went through the code and couldn't find a reason for the inaccurate record. My accuracy generated from the program given was 12.5%. I then went through my data and realized I had marked the Brewers as a American League team, which I changed on the spot. It changed the record generated to 59-3. Progress. I then decided to add a new attribute for teams that have made the playoffs. There was no change to the 59-3 record and I was starting to become worried. I looked over the original data given and realized that they didn't have a row for the actual number of event (ex: game #). I deleted that column completely and the code finally generated a record of 25-37, only one (ONE!!) win away from the 2015 Phillies actual record. I looked over all my data to make sure that it was error-free, and saw that I had set the Mets as a playoff team for 3 games, which I changed and still got a 25-37 record, and a 50% accuracy from the program. I tested all the attributes (by removing them and leaving the other five) to see which had the biggest impact. I listed the record generated WITHOUT that certain attribute, the amount of wins off the actual record, and the program accuracy generated. I then included the winning percentage difference (ex: wp% of day games - wp% of night games). I included this because the higher the difference, the more impact it has at dictating a win or a loss.

Attribute	Record WITHOUT	Wins Off Actual	Program Accuracy	Winning % of Att.
Home/Away	40-22	14 over	45.16%	0.136
Opponent	27-35	1 over	50.00%	
Day/Night	25-37	1 under	50.00%	0.067
League	20-42	6 under	54.84%	0.015
In/Out of Division	25-37	1 under	50.00%	0.030
Playoff Team?	37-25	11 over	46.77%	0.029