

빅데이터를 이용한 프로야구 승부예측

201511005 강우송

201511013 권남호

0. Abstract

빅데이터 분석 및 시각화 개론(SE377) 15조의 연구 주제와 발표 내용을 정리하였다. 주제는 ‘빅데이터를 이용한 프로야구 승부예측’이다. 야구 경기에서의 기록이나 통계와 같은 야구 내부 데이터부터, 프로야구 관중 수, 시즌 당 팀 별 FA 계약 현황부터 google 키워드 검색 기록과 같은 야구 외부 데이터를 2011년부터 2017년까지 수집하였다. 그 후 축적된 데이터를 바탕으로 각 데이터간의 상관관계계수를 결정하였다. 순위를 예측해보려고 하는 시즌에 대해, 직전 시즌의 데이터와 상관관계계수를 행렬 곱셈하여 예상 순위를 산출해 보았다. 이 과정에서 데이터의 선택 여부와 상관관계의 범위에 따른 가중치 부여 등의 다양한 방법을 이용하였으며, 각 방법에 대한 부가적 설명과, 부족했던 부분에 대해 기술한다.

1. Introduction

A. Motivation

다음소프트에서 빅데이터를 활용하여 프로야구 순위를 예측하였다. 그러나 그 예측 결과는 정확하지 않았는데, 우리는 그 이유를 다음과 같이 분석하였다.

1. 사용한 데이터의 폭이 좁다. 다음소프트는 경제 지표, 실력 지표, SNS 지표 등 3가지의 변수만을 사용하였다.
2. 사용한 데이터의 양이 적다. 다음소프트는 지난 2년간의 데이터를 바탕으로 프로야구 순위를 계산하였다.










다음소프트에서 부족했던 부분인 데이터의 폭과 양을 보완하여 진행하면 좀 더 나은 예측 순위를 도출할 수 있을 것이라 생각되어, 좀 더 많고 다양한 데이터를 이용한 순위 예측 알고리즘 제작을 연구 목표로 삼았다.








B. Data

데이터는 많고 다양할수록 좋다고 생각되어, 야구 경기의 결과로 얻어지는 기록과 같은 야구 내부 데이터뿐만 아니라, 야구 경기의 결과 이외의 야구 관련 데이터인 야구 외부 데이터들도 수집하였다. 2011 ~ 2017시즌 사이의 데이터들을 수집하였으며, 팀 시즌 기록(team), 타자

시즌 기록(batter), 투수 시즌 기록(pitcher), 수비 시즌 기록(defense) 등의 야구 내부 데이터는 한국프로야구 통계 사이트인 STATIZ(<http://www.statiz.co.kr/main.php>) 에서의 web-crawling을 통해 추출하였으며, 시즌 관중 수 정보(audience), 선수 연봉 정보(salary), 선수 신체 정보(physical), 구단 FA 상황(FA) 등은 인터넷에서 정보를 모아 취합하였다.

마지막으로 미국의 닭날개 가설을 이용하여 google 키워드 검색을 진행하였다. 닭날개 가설이란, 미국 미식축구리그 NFL에서 닭날개 판매 실적을 조사한 결과 판매 실적이 좋은 도시의 팀이 좋은 성적을 거두었고, 따라서 닭날개 판매 실적과 성적이 상관관계가 있다는 가설이다. 이 가설에 맞게 진행하기 위해서 google에 팀 이름과 치킨이 같이 나오는 키워드 검색을 진행하였다. 다만, 특정 팀의 결과 별명 혹은 팀 마스코트가 치킨으로 불리는 경우가 있기 때문에, (팀 이름 + 치킨 - 팀 별명 - 마스코트)와 같이 검색함으로써 최대한 다른 용어로서 쓰이는 치킨을 배제하려고 노력하였다.

 audience	2017-12-15 오전...	파일 폴더	
 batter	2017-12-15 오전...	파일 폴더	
 defense	2017-12-15 오전...	파일 폴더	
 FA	2017-12-15 오전...	파일 폴더	
 google	2017-12-15 오전...	파일 폴더	
 physical	2017-12-15 오전...	파일 폴더	
 pitcher	2017-12-15 오전...	파일 폴더	
 salary	2017-12-15 오전...	파일 폴더	
 team	2017-12-15 오전...	파일 폴더	

 physical11	2017-12-15 오전...	Microsoft Excel ...	32KB
 physical12	2017-12-15 오전...	Microsoft Excel ...	33KB
 physical13	2017-12-15 오전...	Microsoft Excel ...	33KB
 physical14	2017-12-15 오전...	Microsoft Excel ...	33KB
 physical15	2017-12-15 오전...	Microsoft Excel ...	33KB
 physical16	2017-12-15 오전...	Microsoft Excel ...	33KB
 physical17	2017-12-15 오전...	Microsoft Excel ...	33KB

순위	팀명	평균년차	평균연령	평균신장	평균체중
1	두산	8	26.5	183	85
2	NC	7.5	26.4	183	86
3	넥센	7.1	25.9	182	87
4	LG	8.4	27.5	183.6	85.4
5	KIA	9.2	28.3	183	90
6	SK	8.8	27.7	182	87
7	한화	10.4	29.4	183	87
8	롯데	9	27.9	183.8	90.6
9	삼성	7.8	27.6	181	82.7
10	kt	8.4	27.5	181.7	85.1

2. Methodology

우선 취합한 데이터를 전처리 과정을 거쳐 Null value 등을 제거하였다. 그 후 전처리 과정을 마친 데이터들을 시즌 별로 모아 팀 이름을 기준으로 취합하였다. 시즌 별로 취합한 데이터들을 각 데이터 항목 column과 팀 이름 index에 맞추어 더하고, 취합한 시즌 수로 나누어 모든 데이터가 7년간의 평균을 갖도록 하였다. 이 과정에서 2013년 창단한 NC Dinos와 2015년 창단한 kt wiz의 경우 별도로 분리하여 데이터를 취합하고, 각각 5시즌과 3시즌으로 나누어 균등한 데이터 평균 방식을 취하였다.

7시즌간 취합한 93항목의 데이터에 대하여, pandas의 co-relation 기능을 이용하여 각 데이터마다 상관관계계수를 도출하였다. 따라서 (93 * 93) 크기의 행렬을 얻었고, 그 중 우리는 순위에 가장 영향을 많이 끼치는 요소는 바로 승률이라고 생각하고, 승률에 대한 데이터들의 상관관계계수를 추출하여 분석을 진행하였다. 분석에 관해서는 4. Analyze에서 진행한다.

그 결과, 흥미로운 상관관계계수도 있었지만, 일반적인 야구 상식과 상관관계계수가 대체로 들어맞는 것을 확인할 수 있었고, 이 상관관계계수를 이용하여 시즌마다의 각 팀 성적을 예측하는 방법을 고안하였다.

성적을 예측하려는 시즌을 X라 할 때, 2011 ~ 2017 시즌 데이터를 누적시켜 얻은 상관관계계수 중 승률에 대한 상관관계계수 중 일정 범위에 해당하는 상관관계계수를 뽑은 행렬 C(전체 데이터 선택 시, 크기 93 * 1)와 X-1 시즌 정규화 데이터에서 선택된 상관관계지수에 대응되는 데이터를 선택한 행렬 D_{X-1} (전체 데이터 선택 시, 크기 10 * 93)을 곱하여 X 시즌에 대한 승리 지수 행렬 W_X (크기 10 * 1)는 다음과 같이 계산된다.

$$W_X = D_{X-1} \cdot C$$

이 승리 지수(winning index)를 추출할 수 있을 것이라 생각하였고, 그 승리 지수를 내림차순

으로 정렬하면 그 시즌에 대해 승리 지수가 높은 팀부터 나열될 것이고, 이것이 곧 우리가 예측한 그 시즌의 성적이다. 상관관계계수에 범위를 지정한 이유는 4. Analyze에서 설명한다.

3. Result

2011 ~ 2017 시즌 데이터를 누적하여 계산한 상관관계계수 중 승률과의 상관관계이다.


	승률
순위	-0.994617
승	0.995568
패	-0.992332
무	0.718459
승률	1.000000
게임차	-0.996331
batter_2타	0.209421
batter_3타	0.836797
batter_G	-0.617239
batter_OPS	0.825165
batter_WAR*	0.915354
batter_WPA	0.785555
batter_wOBA	0.884473
batter_wRC+	0.948799
batter_고4	0.388299
batter_도루	0.702163
batter_도실	-0.078127
batter_득점	0.740852
batter_루타	0.419992
batter_병살	-0.523041
batter_볼넷	0.510596
batter_사구	0.647503
batter_삼진	-0.567884
batter_안타	0.075206
batter_장타	0.726055
batter_출루	0.837555
batter_타석	-0.382905
batter_타수	-0.502059
batter_타율	0.687345
batter_타점	0.717227
batter_홈런	0.423244
batter_희비	0.498169
batter_희타	-0.151842

pitcher_2타	-0.685704
pitcher_3타	-0.745708
pitcher_ERA	-0.918680
pitcher_ERA+	0.862739
pitcher_FIP	-0.759710
pitcher_FIP+	0.664739
pitcher_WAR	0.925068
pitcher_WHIP	-0.881304
pitcher_WPA	0.890181
pitcher_고4	-0.273770
pitcher_보크	-0.404892
pitcher_볼넷	-0.666168
pitcher_사구	-0.452775
pitcher_삼진	0.065695
pitcher_세	0.873629
pitcher_승	0.995575
pitcher_실점	-0.931157
pitcher_안타	-0.856082
pitcher_완봉	-0.117931
pitcher_완투	-0.121916
pitcher_이닝	-0.588383
pitcher_자책	-0.917027
pitcher_출장	-0.798105
pitcher_타자	-0.865338
pitcher_패	-0.992332
pitcher_폭투	-0.845863
pitcher_홀드	0.688620
pitcher_홈런	-0.533264
defense_/133	0.895144
defense_ARM	0.614615
defense_BLK	0.866985
defense_CS	0.526458
defense_E+	0.876413
defense_POSADJ	-0.660767
defense_RAA	0.893342
defense_RAAwithADJ	0.767675
defense_RF9	-0.588144
defense_RNG	0.814485
defense_WAAw/oADJ	0.890512
defense_WAAwithADJ	0.748261
defense_기회	-0.600660
defense_보살	-0.642587
defense_선발	-0.583960
defense_수비율	-0.560463
defense_실책	-0.769982

defense_이닝	-0.574589
defense_자살	-0.574072
defense_출장	-0.226637
audience_avg	0.201255
audience_total	0.166027
salary_avg	0.407805
salary_total	0.408621
physical_평균년차	-0.205040
physical_평균신장	-0.109572
physical_평균연령	-0.207300
physical_평균체중	-0.281222
FA_FA수익	0.532095
FA_FA지출	-0.198543
FA_순수익	0.615028
google_검색	0.304994

2011 ~ 2017 시즌 데이터를 누적하여 계산한 상관관계계수와 2016시즌 정규화 데이터를 이용하여 2017년 시즌 성적을 예측해 보았다. 다음은 실제 2017시즌 한국프로야구 순위이다.

2017 정규리그순위

순위	팀명	경기	승	무	패	승률	게임차	연속	최근10경기
1	 KIA	144	87	1	56	0.608	0.0	2승	6승-4패-0무
2	 두산	144	84	3	57	0.596	2.0	1패	8승-2패-0무
3	 롯데	144	80	2	62	0.563	6.5	5승	8승-2패-0무
4	 NC	144	79	3	62	0.560	7.0	4승	5승-4패-1무
5	 SK	144	75	1	68	0.524	12.0	2승	7승-3패-0무
6	 LG	144	69	3	72	0.489	17.0	2패	4승-6패-0무
7	 넥센	144	69	2	73	0.486	17.5	4패	3승-7패-0무
8	 한화	144	61	2	81	0.430	25.5	5패	3승-6패-1무
9	 삼성	144	55	5	84	0.396	30.0	2승	4승-6패-0무
10	 kt	144	50	0	94	0.347	37.5	2패	3승-7패-0무

승리 지수를 예측하는 과정에서 다음과 같은 세 가지 모델을 사용하였다.

1. 데이터의 일부만 사용: 야구 외부 데이터만을 사용한 순위 예측
2. 전체 데이터 사용: 야구 내부, 외부 데이터를 모두 사용한 순위 예측

3. 상관관계계수의 범위에 따른 가중치 부여: 야구 내부, 외부 데이터와 가중치를 사용한 순위 예측

1. 데이터의 일부만 사용

혹시 야구 외부 데이터가 야구에 얼마나 영향을 끼치는 지 조사하기 위하여 야구 내부 데이터를 모두 배제하고 순위를 예측해 보았다. 그 결과는 다음과 같다.

	winning index
팀명	
SK	1.000000
삼성	0.924536
한화	0.431702
두산	0.329601
LG	0.323934
넥센	0.296434
롯데	0.278396
KIA	0.263894
NC	0.143874
kt	0.000000

2. 전체 데이터 사용

모든 데이터를 이용하여 예측한 2017시즌 프로야구 성적은 다음과 같다.

	winning index
팀명	
두산	1.000000
NC	0.965258
SK	0.838700
삼성	0.725093
넥센	0.678004
LG	0.568811
KIA	0.542118
롯데	0.489461
한화	0.383908
kt	0.000000

3. 상관관계계수의 범위에 따른 가중치 부여

다음은 일종의 Manual-Tuning을 이용한 상관관계계수의 범위와, 그에 따른 가중치이다.

상관관계계수 범위
$t1 = t1[t1['승률'] > 0.3]$
$t1 = t1[t1['승률'] < 0.4]$
$t2 = t2[t2['승률'] \geq 0.4]$
$t2 = t2[t2['승률'] < 0.5]$
$t3 = t3[t3['승률'] \geq 0.5]$
$t3 = t3[t3['승률'] < 0.65]$
$t4 = t4[t4['승률'] \geq 0.65]$
$t4 = t4[t4['승률'] < 0.8]$
$t5 = t5[t5['승률'] \leq -0.3]$
$t5 = t5[t5['승률'] > -0.4]$
$t6 = t6[t6['승률'] \leq -0.4]$
$t6 = t6[t6['승률'] > -0.5]$
$t7 = t7[t7['승률'] \leq -0.5]$
$t7 = t7[t7['승률'] > -0.65]$
$t8 = t8[t8['승률'] \leq -0.65]$
$t8 = t8[t8['승률'] > -0.8]$
가중치 부여
$t1 = t1 * 1$
$t2 = t2 * 2$
$t3 = t3 * 2.5$
$t4 = t4 * 3$
$t5 = t5 * 0.9$
$t6 = t6 * 1.8$
$t7 = t7 * 2.2$
$t8 = t8 * 2.7$

이 가중치를 적용했을 때의 순위 예측은 다음과 같다.

	winning index
팀명	
두산	1.000000
NC	0.990986
넥센	0.949292
삼성	0.630713
KIA	0.606054
SK	0.555828
LG	0.540548
롯데	0.507848
한화	0.195963
kt	0.000000

4. Analyze

우선 첫 번째 모델에서 볼 수 있듯 예측 결과는 실제 예측 순위와 많이 다르며, 야구 외부 데이터는 실제 야구단의 승률에 많은 영향을 끼치지 못한다고 해석할 수 있다.

반면 두 번째 모델, 내부와 외부 데이터를 모두 사용했을 때의 예측 결과는, 완벽하게 결과를 예측해내지는 못했지만, 상위권, 중위권, 하위권으로 크게 카테고리를 나눠 보았을 때는 흥미롭게 예측할 수 있었다.

세 번째 모델 역시 두 번째 모델과 비슷한 양상을 나타냈으며, 특히 하위권에 위치한 kt는 아직 창단 3년차인 신생 팀인 관계로 모든 분석에서 10위 자리를 피할 수 없었다. 이는 kt 구단이 승률과 관련된 대부분의 지표에서 다른 팀들에게 밀린다고 해석할 수 있다.

상관관계계수가 너무 극단적인 경우는 너무 당연한 데이터들이 포함되는 것을 발견하였다. 예를 들어 팀의 승리 횟수가 승률에 99% 정도의 상관관계를 보이는데, 이는 승률이 승리에 의해 계산되는 점을 생각하면 당연한 것이며, 따라서 70% 이상의 상관관계를 기준으로 데이터를 제한하였다.

각 모델 모두 상관관계계수 a 가 $0.4 < a < 0.7$, $-0.7 < a < -0.4$ 인 경우만 추출하여 분석을 진행하였다.

5. Discussion

실제 결과와 비슷한 결과를 얻었지만, 상관관계계수는 7시즌 정도의 데이터에서는 그 변동 폭이 매우 크며, 통계적인 검증도 거치지 않았기 때문에 결과를 함부로 신뢰하기 힘들다. 신뢰하기 힘든 상관관계계수와 시즌 정규화 데이터를 행렬 곱셈을 하는 것이 유효한 결과를 낼 수

있는가에 대해서는 통계적 검증을 더욱 철저하게 마쳐야 한다.

한편 별개로 더 많은 일정한 관계가 있는 데이터(야구는 데이터가 쌓일수록 특정 결과에 수렴하는 양상을 많이 보인다)를 축적하면 상관관계계수가 안정화되어 통계적으로 검증하기도 더욱 쉬워질 것이다.

모델 3에서는 가중치를 일종의 manual-tuning을 이용하여 계산하였는데, PCA를 통한 주성분 분석으로 관련 변수들을 묶어 93개의 데이터를 수 개 수준으로 줄일 수 있으며, 이를 Linear Regression 등을 통하여 좀 더 좋은 가중치를 계산할 수 있을 것이라 생각된다.