

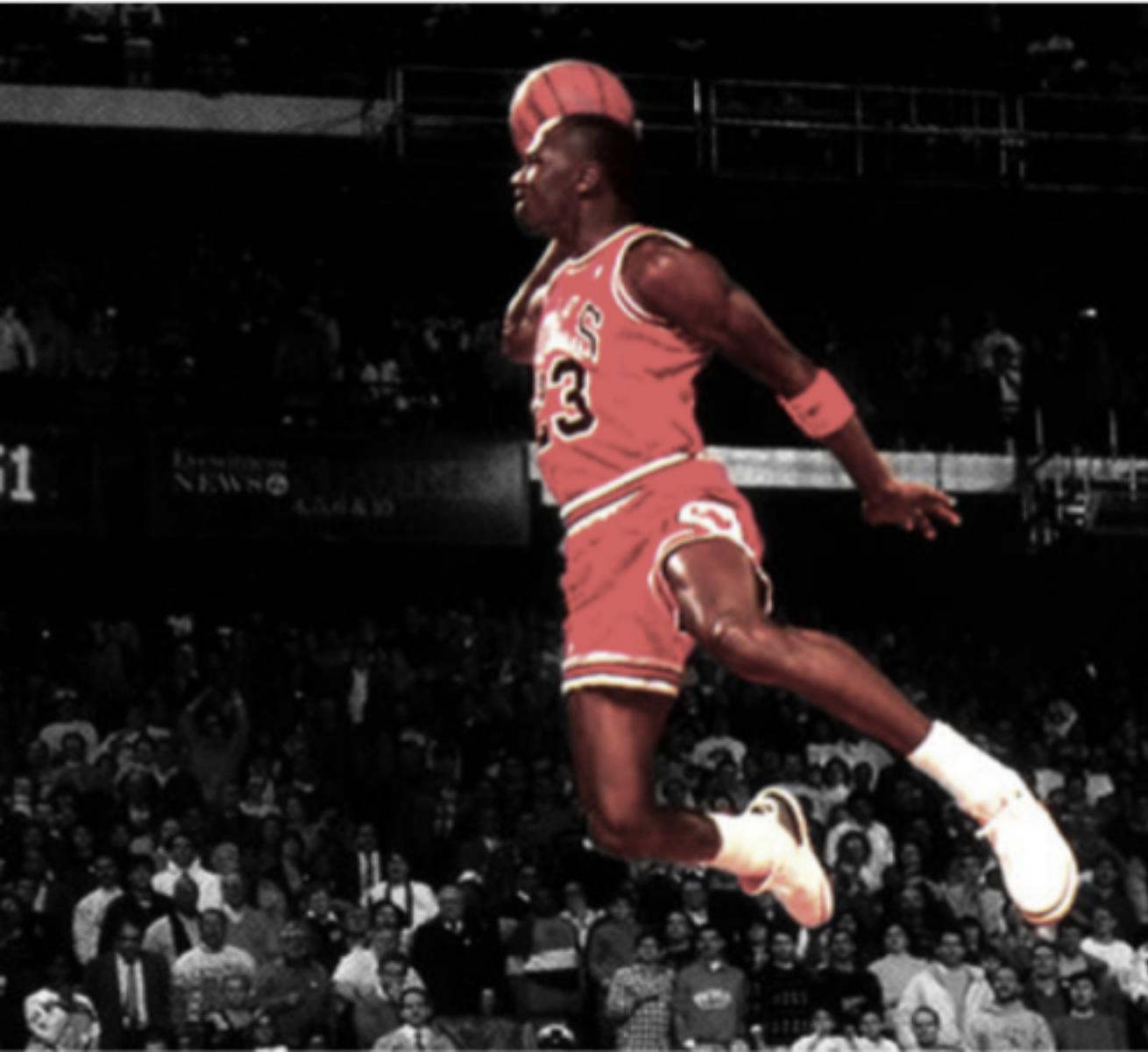
NBA 데이터 분석

classification을 통한 현액 가능성 예측 &
regression 통한 salary 예측을 중심으로

6조

임 규 민 | 강 희 우 | 장 병 인 | 이 해 나





INDEX

01 주제 소개

선정 이유, 중요성

02 데이터 소개

데이터 수집 방법, 데이터 사이즈, 변수 설명

03 전처리

적용한 전처리 방법, 적용 이유

04 분석 방법

변수 선택 방법, 적용 알고리즘, 평가, 결과해석

05 새로운 예측값-1

Salary ratio

06 새로운 예측값-2

Salary

07 결론 및 시사점

시사점 및 프로젝트 후기

01 주제 소개

선정 이유, 중요성

- 현역 선수들의 명예의 전당 입성과 관련된 수 많은 논쟁들이 존재
- 기존의 명예의 전당 헌액 선정 방식의 한계 - 100% 주관 (voting system)

이제 달라 명예의 전당 가능할까요?

KD James 2017-12-23 14:59:18 246

루커킹스터트, 폴스터1회, 디펜시브밀2회(파스로 부네요... 전설기록을 NBA에 없는게 아쉽지만...) 끝연수인줄 몰랐던 입성을 가능하게 하셨습니다? 몽전 같았던 그 외 몇개를 개인적으로 느꼈을 때 그렇게 일찍트가 있는 MVP는 아니였다고 생각이 들거든요. 스트로보원 슈퍼스타급 스템포도 아니였고...

스프링 보자면 26-7-4 오프 49% 안되고 3점 35% 가 단 합니다. 90%로 이후 매우 40% 안되는 MVP는 꽤 2점 단에 아티스트, 서비석, 포즈 좋아해요. 근데 당시는 대로 2점은 정체이 30점을 넘긴 선수들이죠. 95-96 폐나탈 비교해도 훨씬 더 열어지는 편이고, 물론 시즌같은 패스나가 있어서 이런 단위로 그 시즌 열심히 경기했으므로서도 그렇고 스트론으로도 그렇고 MVP급 선수로서 일찍트한 거라고는 하면 일찍트가 대단했다 이 정도는 아닌거 같습니다.

오히려 폴лен도 시즌 학번이나 학년 서비석을 보고 일찍트가 유행났다 어려운 말이 둘 되는데 폴은 MVP 시즌이 오후스터급에선 일찍트가 둘 뒤 유행이라고 생각해버려요.

다른분들은 어떻게 생각하시는지 궁금하네요. 일찍트라는 부분에서 내부도 둘 그립간판에 내쉬는 게임 조립 능력 자체가 폴즈보다는 한 우상계 뒷 글이라고 생각되네요.

13 Comments

11 Comments



01 주제 소개

선정 이유, 중요성

농구 관련 커뮤니티에서 항상 논란이 되는 것 :

이 선수가 은퇴 후 명예의 전당에 들어갈 수 있는가?

**정성적 평가가 아닌 정량적 평가기준(경기 내 스탯 등)을 더 도입해
명예의 전당 입성 여부를 평가해보기로 함.**

02 데이터 소개

데이터 수집 방법, 데이터 사이즈, 변수 설명

From Kaggle

Player data

이름, 키, 몸무게, 대학, 출생년도 등이 적혀있음.

은퇴년도(Year_end)를 기준으로 1980~ 2010년에 은퇴한 선수들만 선택.

name	year_start	year_end	position	height	weight	birth_date	college		
Alaa Abd	1991	1995	F-C	06월 10일	240	24-Jun-68	Duke University		
Zaid Abd	1969	1978	C-F	06월 09일	235	07-Apr-46	Iowa State University		
Kareem Al	1970	1989	C	07월 02일	225	16-Apr-47	University of California, Los Angeles		
Mahmoud	1991	2001	G	06월 01일	162	09-Mar-69	Louisiana State University		
Tariq Abd	1998	2003	F	06월 06일	223	03-Nov-74	San Jose State University		
Shareef Al	1997	2008	F	06월 09일	225	11-Dec-76	University of California		
Tom Aber	1977	1981	F	06월 07일	220	06-May-54	Indiana University		
Forest Abl	1957	1957	G	06월 03일	180	27-Jul-32	Western Kentucky University		
John Abr	1947	1948	F	06월 03일	195	09-Feb-19	Salem International University		
Alex Abrin	2017	2018	G-F	06월 06일	190	01-Aug-93			
Alex Acke	2006	2009	G	06월 05일	185	21-Jan-83	Pepperdine University		
Don Acke	1954	1954	G	6-0	183	04-Sep-30	Long Island University		
Mark Acre	1988	1993	F-C	06월 11일	220	15-Nov-62	Oral Roberts University		
Bud Actor	1968	1968	F	06월 06일	210	11-Jan-42	Hillsdale College		
Quincy Ac	2013	2018	F	06월 07일	240	06-Oct-90	Baylor University		

02 데이터 소개

데이터 수집 방법, 데이터 사이즈, 변수 설명

From Kaggle

Season data

선수 이름 별로 경기 내 스탯들이 적혀있음.

Data Shape : 24692(가로)*54(세로)

수많은 Feature들 중, 원하는 raw data 몇 가지와
유용한 2차 스탯들만 사용하기로 하였다.

3	1950 Ed Daniels F	24 TOT	15	0.312	0.395	-0.5	-0.1	-0.6	22	36	0.256	0.256	19	34	0.358	20	62
4	1950 Ed Daniels F	24 DYN	13	0.308	0.378	0	0	0	21	35	0.256	0.256	17	31	0.348	27	63
5	1950 Ed Daniels F	24 NYK	2	0.376	0.75	0	0	0	1	4	0.25	0.25	2	2	0.657	2	4
6	1950 Ralph Saw G	22 INO	65	0.422	0.801	0.16	1.2	0.45	242	256	0.263	0.263	215	282	0.763	132	595
7	1950 Game Score G/H	23 TRI	2	0.373	0.313	-0.1	-0.1	-0.1	5	16	0.213	0.213	0	5	0	6	15
8	1950 Carl Erskine R	25 TOT	25	0.345	0.391	-0.2	-0.5	-0.2	226	247	0.278	0.278	204	221	0.617	277	661
9	1950 Carl Erskine R	25 FTW	26	0.302	0.448	-0.7	-2.1	-1.5	126	143	0.287	0.287	123	209	0.621	140	392
10	1950 Charlie Gehrke F	25 AND	29	0.328	0.394	-1.5	2.8	1.1	121	178	0.287	0.287	71	112	0.688	133	279
11	1950 Nelson E/R G	25 PWK	37	0.398	0.332	0.4	1.2	1.8	80	248	0.222	0.222	82	133	0.626	87	242
12	1950 Jake Scott F/C	22 PWK	60	0.358	0.394	-0.7	1.5	0.5	55	205	0.289	0.289	73	117	0.687	40	111
13	1950 Vince Sisko SF	22 NYK	19	0.426	0.445	0.8	1.4	1.9	204	802	0.34	0.34	204	287	0.764	203	612
14	1950 Don Sisko F/G	24 WAT	62	0.461	0.623	0.44	-0.7	1.6	208	158	0.273	0.273	240	249	0.688	235	598
15	1950 Harry Sisko C	27 WAT	61	0.479	0.375	0	-0.7	1.2	208	685	0.413	0.413	203	282	0.775	229	779
16	1950 Jim Brown C	21 DYN	46	0.347	0.347	-0.1	-0.7	-0.2	26	147	0.269	0.269	15	15	0.671	225	713
17	1950 Jim Brown C	24 DYN	29	0.402	0.494	21	0.7	1.6	254	713	0.228	0.228	245	255	0.671	229	713
18	1950 Carl Erskine R	22 NYK	67	0.454	0.385	0.2	1.9	7.2	272	1204	0.264	0.264	285	274	0.782	147	1201
19	1950 Frankie Shi G	26 AND	64	0.415	0.422	0.6	1.6	9.2	268	1158	0.215	0.215	402	485	0.824	182	1128
20	1950 Fredrick G/C	29 RDC	7	0.392	0.581	0.05	0.1	0.5	11	22	0.478	0.478	12	12	0.822	7	24
21	1950 Bob Brown F	26 DYN	52	0.414	0.322	1.1	-0.9	1.2	278	764	0.261	0.261	172	255	0.883	101	724
22	1950 Jim Brown C	20 DYN	21	0.392	0.583	0.01	-0.1	0	17	48	0.254	0.254	12	27	0.451	16	47
23	1950 Walt Bell G	24 SLB	68	0.385	0.403	0.7	2	2.7	198	652	0.204	0.204	199	282	0.757	195	595
24	1950 Jack Dunn G	23 SHG	61	0.378	0.283	0.1	-0.4	-0.4	227	711	0.222	0.222	124	181	0.881	179	527
25	1950 Jim Brown C	24 DYN	42	0.368	0.313	0.2	-0.4	-0.4	129	247	0.260	0.260	129	184	0.881	149	527
26	1950 Ed Calfee G/S	23 RDC	62	0.449	0.371	0.3	1.5	4.5	207	548	0.277	0.277	146	203	0.718	113	565
27	1950 Don Carter G/C	20 MNL	57	0.402	0.322	0.7	1.9	2.8	98	292	0.241	0.241	69	95	0.726	128	267
28	1950 Bob Carty F	22 FTW	58	0.401	0.401	0.2	2.8	5	212	617	0.244	0.244	192	255	0.742	168	614
29	1950 Jake Carney F/C	25 TOT	34	0.417	0.707	0.3	0.4	0.7	22	75	0.207	0.207	36	53	0.678	39	52
30	1950 Jake Carney F/C	25 DYN	13	0.39	0.578	0.01	-0.1	0	13	48	0.299	0.299	18	26	0.692	27	44
31	1950 Jake Carney F/C	25 AND	11	0.454	0.39	0.2	0.3	0.7	15	35	0.223	0.223	15	27	0.687	22	55
32	1950 Jim Carter F	22 SHG	55	0.347	0.385	1.9	2.4	8.2	142	421	0.232	0.232	207	246	0.727	227	572
33	1950 Jim Carter F	22 SHG	52	0.354	0.327	0.02	-0.1	-0.1	28	292	0.224	0.224	20	29	0.625	22	52
34	1950 John Chen G/C	29 TBL	6	0.331	0.354	0.01	0.1	0.5	15	17	0.277	0.277	5	12	0.687	15	28
35	1950 John Chen G/C	29 SHG	10	0.372	0.347	0	0	-0.1	15	49	0.206	0.206	12	17	0.708	5	42
36	1950 Leroy Chon F	24 SVR	49	0.385	0.313	0.01	0.8	0.9	61	179	0.241	0.241	25	58	0.625	12	157
37	1950 Bill Cross SF	28 AND	64	0.372	0.326	-0.5	4.9	4.4	282	658	0.215	0.215	198	288	0.715	180	702
38	1950 Paul Croyd G/H	29 TOT	7	0.322	0.382	-0.1	0	-0.1	7	26	0.269	0.269	5	8	0.678	2	15
39	1950 Paul Croyd G/H	29 SLB	3	0.368	0.375	0.01	0	-0.1	1	5	0.125	0.125	3	2	1	4	5
40	1950 Jack Dunn G	22 SHG	4	0.347	0.371	0.01	0	0	6	12	0.222	0.222	2	4	0.647	1	14
41	1950 Jack Dunn G	22 RDC	65	0.412	0.585	1.9	2.2	4.4	280	658	0.277	0.277	98	121	0.624	229	592
42	1950 Bobby Crotty F	24 SHG	51	0.42	0.392	21	-0.1	1.9	222	620	0.258	0.258	143	181	0.79	158	527
43	1950 Ray Corde G	22 SVR	62	0.365	0.333	0.02	1.6	1.2	117	272	0.216	0.216	78	122	0.615	81	209
44	1950 Jack Crotty F/C	25 DYN	54	0.343	0.403	-0.1	-0.5	-1.5	27	322	0.222	0.222	92	161	0.509	65	154
45	1950					-0.1	-0.1	-1.5	27	322	0.222	0.222	92	161	0.509	65	154

02 데이터 소개

데이터 수집 방법, 데이터 사이즈, 변수 설명

By 매크로 사용

Search data

선수 이름 별로 Google 검색을 이용해서 검색 수를 수합.

Google Gus Bailey nba

전체 이미지 뉴스 동영상 지도 디보기

검색결과 약 363,000개 (0.46초)

Gus Bailey Stats | Basketball-Reference.com
https://www.basketball-reference.com/Players/B/Gus_Bailey.html

Gus Bailey - Wikipedia
https://en.wikipedia.org/wiki/Gus_Bailey

NBA.com/Stats | Gus Bailey
https://stats.nba.com/players/134/

거스 베일리

농구 선수

출생: 1951년 2월 18일, 미국 노스캐롤라이나주 갑슨
사망 날짜/장소: 1988년 11월 28일, 미국 루이지애나 주 뉴올리언스
키: 196cm
체중: 84kg
학력: 텍사스 대학 엘파소
NBA 드래프트: 1974년
포지션: 스몰 포워드, 슈팅 가드

name	search_num
Gus Bailey	694,000
Marvin Barnes	804,000
Rick Barry	4,330,000
Tim Bassett	314,000
Ron Behagen	11,700
Del Beshore	4,320
Lawrence Boston	7,530,000
Alonzo Bradley	346,000
John Brown	116,000,000
Roger Brown	29,600,000
Corky Calhoun	18,200
Bob Carrington	396,000
Ron Carter	6,310,000
Don Chaney	294,000
Jim Cleamons	67,500
John Coughran	9,420
Terry Crosby	575,000
Harry Davis	12,200,000

02 데이터 소개

데이터 수집 방법, 데이터 사이즈, 변수 설명

By 직접 수집

Hall of Fame data

명예의 전당에 들어간 선수들의 데이터를 모두 모아
이름 별로 0과 1로 표기함.

name	H_O_F
Paul Silas	0
Rick Barry	1
Walt Frazier	1
Phil Jackson	0
Earl Monroe	1
Don Chaney	0
Ron Boone	0
Wes Unseld	1
Elvin Hayes	1
Bingo Smith	0
Mack Calvin	0
Jo Jo White	1
Bob Dandridge	0
Spencer Haywood	1
Steve Mix	0
Kareem Abdul-Jabbar	1

02 데이터 소개

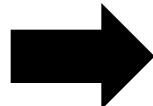
데이터 수집 방법, 데이터 사이즈, 변수 설명

By 직접 수집

우승 횟수 data

1965년부터 2010년 까지 모든 우승팀의 로스터를 모아,
이름 별로 정렬한 뒤, 선수 별로 정리해주었다.

SG	21		Bonham, Ron	6 ft 3 in (1.96 m) 32 lb (87 kg)	Cincinnati
C	11		Counts, Mel	6 ft 0 in (2.13 m) 30 lb (104 kg)	Oregon State
G/F	28		Green, Si	6 ft 2 in (1.88 m) 35 lb (84 kg)	Duquesne
G/F	17		Havlicek, John	5 ft 5 in (1.96 m) 33 lb (92 kg)	Ohio State
PG	25		Jones, K. C.	6 ft 1 in (1.85 m) 30 lb (91 kg)	San Francisco
SG	24		Jones, Sam	6 ft 1 in (1.93 m) 38 lb (90 kg)	North Carolina Central
SF	12		Naulls, Willie	6 ft 5 in (1.98 m) 55 lb (102 kg)	UCLA
F	19		Nelson, Don	6 ft 5 in (1.98 m) 10 lb (95 kg)	Iowa
C	6		Russell, Bill	6 ft 9 in (2.06 m) 15 lb (98 kg)	San Francisco
SF	16		Sanders, Satch	6 ft 5 in (1.98 m) 10 lb (95 kg)	New York



name	Final win
Paul Silas	3
Rick Barry	1
Walt Frazier	1
Phil Jackson	1
Earl Monroe	1
Don Chaney	2
Ron Boone	0
Wes Unseld	1
Elvin Hayes	1
Bingo Smith	0
Mack Calvin	0
Jo Jo White	2
Bob Dandridge	2
Spencer Haywood	1
Steve Mix	0
Kareem Abdul-Jabbar	6

03 전처리

적용한 전처리 방법, 적용 이유

Merge

- 앞서 나온 데이터들의 shape이 모두 달랐기에, 1904명의 선수별 name을 기준으로 merge 해야했음
- 하지만 단순 merge를 했을 때 자꾸 총 row 수가 1904와 달라지는 문제가 지속적으로 발생
- 그 원인은 **(1)동명이인 (2)별표 붙은 이름 (3)season_stat_data에서 3단어 이상의 선수 이름 누락**

1

```
[183]: runfile('/Users/hanameee/.spyder-py3/temp.py', wdir='/Users/hanameee/.spyder-py3')
    출복된 Bobby Jones 이 발견되었습니다
    출복된 Cedric Henderson 이 발견되었습니다
    출복된 Charles Jones 이 발견되었습니다
    출복된 Charles Jones 이 발견되었습니다
    출복된 Charles Smith 이 발견되었습니다
    출복된 Charles Smith 이 발견되었습니다
    출복된 Dee Brown 이 발견되었습니다
    출복된 Eddie Johnson 이 발견되었습니다
    출복된 George Johnson 이 발견되었습니다
    출복된 Ken Johnson 이 발견되었습니다
    출복된 Marcus Williams 이 발견되었습니다
    출복된 Mark Davis 이 발견되었습니다
    출복된 Mark Jones 이 발견되었습니다
    출복된 Michael Smith 이 발견되었습니다
```

2

1965	John Barnhill	S
1965	Elgin Bay*pr*	S
1965	Zelmo Beaty*	C

3

1971	Jo Jo
1970	Jo Jo White

→ 동명이인 삭제, 별표 삭제, 3단어 이상 이름 누락된 것을 수정하고 앞서 나온 데이터들을 merge 하였음

03 전처리

적용한 전처리 방법, 적용 이유

Season Stat을 2가지로 분류

- 선수를 평가하기 위해서는 한 시즌의 **단기 임팩트(season-high)**와 꾸준함을 나타내는 **스탯 합(sum stat)**을 모두 고려해야 함
- 이를 위해서 Season Stat 데이터에서 선수별 groupby를 통한 시즌 Max stat을 구했고, 선수별로 stat을 모두 더해 sum stat을 따로 구했음. 이후 이 Max(season-high)와 Sum(선수 생활 전체 스탯 합)을 각각 다른 컬럼으로 해서 추가했음

Max Stat						
name	G	PER	WS	TRB	AST	PTS
A.C. Green	1361	249.7	104.20000000000002	10129	1469	12928
A.J. Bramlett	8	-0.4	-0.2	22	0	8
A.J. English	151	23.1	1.1	315	320	1502

Sum stat						
G_sum	PER_sum	WS_sum	TRB_sum	AST_sum	PTS_sum	
1361	249.7	104.20000000000002	10129	1469	12928	
8	-0.4	-0.2	22	0	8	

03 전처리

적용한 전처리 방법, 적용 이유

NAN 값 처리

- 표준화를 하기 전 PER의 **NAN** 값을 삭제
- 검색 결과 수치의 **comma (,)** 가 정규화 시 에러를 일으켜 comma도 모두 삭제해 줌



표준화

- 변수들 간 데이터 값의 단위가 모두 다르고, 값의 차이도 크기에 **데이터 정규화**를 실시해주었음
- 숫자형 데이터의 분포를 평균 0, 분산 1인 **표준정규분포**로 변환함

search_num	HoF	final_win_num	G	PER
48800000	0	3	83	17.8
11600	0	0	8	-0.4
6870000	0	0	81	12.5
38800	0	0	45	11.8
2910	0	0	6	7.6

search_num	final_win_num	G
2.856597199	4.194524495	0.947490731
-0.361072611	-0.299950782	-1.669409664
0.091249388	-0.299950782	0.87770672
-0.359278729	-0.299950782	-0.378405469
-0.36164573	-0.299950782	-1.739193675

03 전처리

적용한 전처리 방법, 적용 이유

One Hot Encoding

- 수치형 데이터가 아닌 포지션 값들 (POS)에 대해선 one-hot encoding을 실시함

Before		After									
Pos		Pos_C	Pos_PF	Pos_PF-C	Pos_PG	Pos_PG-SG	Pos_SF	Pos_SF-PF	Pos_SG	Pos_SG-SF	HoF
PF		0	1	0	0	0	0	0	0	0	0
C		1	0	0	0	0	0	0	0	0	0
SG		0	0	0	0	0	0	0	1	0	0

04 분석 방법

변수 선택 방법, 적용 알고리즘, 평가, 결과 해석

Decision Tree

- Y값이 Hall of Fame 입성 유무 0,1 두 가지로 나뉘는 Classification 데이터 → **decision tree**를 실시
- 학습 후 score 함수로 accuracy를 확인한 결과, 학습 정확도가 1.0, 검증 데이터 정확도가 0.97 정도의 수치를 보였음

```
In [397]: tree.score(X_train,y_train)  
Out[397]: 1.0
```

```
In [398]: tree.score(X_test,y_test)  
Out[398]: 0.9698581560283688
```

→ 하지만 우리가 가진 HoF 데이터는 0과 1의 비율이 97:3 정도로 **class의 불균형**이 심하기에 **accuracy만으로는 정확도가 왜곡될 염려가 있음.** 따라서 추가적인 분석이 필요하다고 판단했다.

0	1821
1	57

04 분석 방법

변수 선택 방법, 적용 알고리즘, 평가, 결과 해석

분류 성능 평가 - Confusion Matrix & F-score

- **Accuracy** (전체 중에 얼마나 정확하게 맞췄는가?) 외에 **Precision**(예측한 것 중에 얼마나 맞췄는가?) 과 **Recall** (찾아야 하는 것중에 얼마나 찾았는가?) 을 보여주는 **Confusion Matrix**를 사용함. 서로 상충되는 Precision과 Recall을 고려하는 F-score 값 또한 사용함.
- 3%에 불과한 1 데이터를 거의 예측하지 못하는 결과가 나와, Accuracy를 제외한 다른 수치들이 비교적 낮음을 알 수 있었음

```
In [411]: runfile('/Users/hanameee/.spyder-py3/temp.py', wdir='/Users/hanameee/.spyder-py3')
[[542  2]
 [ 20  0]]
accuracy: 0.9609929078014184
precision: 0.4822064056939502
recall: 0.49816176470588236
f1_score: 0.4900542495479204
Traceback (most recent call last):
```

		예측값	
		0	1
실제값	0	542	2
	1	20	0

→ 0과 1 사이의 데이터 불균형으로 인해 문제가 생각한다고 판단, hall of fame 1값을 가진 데이터를 oversampling 해서 다시 모델을 돌려보기로 결정함

04 분석 방법

변수 선택 방법, 적용 알고리즘, 평가, 결과 해석

Oversampling

- 데이터 불균형문제를 해결하기 위해 **oversampling** 을 사용했다. Undersampling 시 전체 데이터 갯수가 너무 적어질 우려가 있어 oversampling 을 사용했으며, 0의 갯수 만큼 1의 갯수를 oversampling했다. (57개 → 1821개)

	precision	recall	f1-score	support	
0	1.00	0.98	0.99	549	
1	0.98	1.00	0.99	544	
avg / total	0.99	0.99	0.99	1093	
accuracy:	0.989935956084172				
precision:	0.9900900900900901				
recall:	0.9899817850637522				
f1_score:	0.9899356528013674				

		예측값	
		0	1
실제값	0	538	11
	1	0	544

→ Oversampling 이후 훨씬 **개선된 결과치**를 얻을 수 있었으나, oversampling은 정확한 실제 데이터가 아니므로 이 주제로 결론을 내리기는 무리가 있다고 판단, **추가 주제로 데이터마이닝을 실시**했음.

05 새로운 예측값 선정 Salary Ratio

선정이유

1. Hall of Fame 데이터의 한계

현액된 선수의 수가 그렇지 못한 선수에 비해 작다 (약 9:1)

2.. 유의미한 타겟

Hall of fame 이 가지는 의미가 드러나는 타겟

추가 데이터 수집

1. 선수의 활동 년도 별 연봉액
수집 후 전처리 예정
 - : <https://www.basketball-reference.com/contracts/players.html>
 - <https://hoopshype.com/salaries/players/>
2. 기존 변수들을 년도별로 다시 수집
 - Season statistics, mvp, all-star, Team grade

05 Salary Ratio

데이터 전처리 과정

A	B		C	D	E
1	YearEnd	Team	Player	Salary	BelowMin
2	2017	Atlanta Hawks	Dwight Howard	#####	
3	2017	Atlanta Hawks	Paul Millsap	#####	
4	2017	Atlanta Hawks	Kent Bazemore	#####	
5	2017	Atlanta Hawks	Tiago Splitter	8,550,000	
6	2017	Atlanta Hawks	Kyle Korver	5,239,437	
7	2017	Atlanta Hawks	Kris Humphries	4,000,000	
8	2017	Atlanta Hawks	Thabo Sefolosha	3,850,000	

전처리 과정

1. 선수의 활동 연도 별 연봉액

1) 데이터 축소

salary data set에서 필요없는 변수 column 지운다

2) 데이터 클리닝

0이 있는 row를 다 없앤다 > salary data에는 0인 element

가 없기 때문에 사용 가능

3) 데이터 변환

player 별 연봉 상승률을 for문으로 구하여 column addition 한다

player 기준으로 groupby.mean() 한다

A	B		C	D	E	F
1	Player	YearEnd	Team	Salary	n_salary	ratio
2	A.C. Green	1991	Los Angeles Lakers	1,750,000	1750000	0
3	A.C. Green	1992	Los Angeles Lakers	1,750,000	1750000	0
4	A.C. Green	1993	Los Angeles Lakers	1,750,000	1750000	0.07714286
5	A.C. Green	1994	Phoenix Suns	1,885,000	1885000	2.43374005
6	A.C. Green	1995	Phoenix Suns	6,472,600	6472600	6.18E-05

05 Salary Ratio

데이터 전처리 과정

전처리 과정

2. 선수의 활동 년도 별 연봉인상률과 기존의 데이터 통합

데이터 통합

salary ratio_mean과 Hof (이미 앞에서 전처리한 데이터)를 merge 한다.

#추가적으로 height 등의 데이터를 merge한다

3. 데이터를 표준화한다.

데이터 정규화 - Standardization

변수들의 분포 차이가 크므로 정규화 시킨다. Standardscaler

05 Salary Ratio

데이터 전처리 과정

A	B	C	D	E	F	G	H	I	J	
1	height	search_num	final_win_num	G_sum	PER_sum	WS_sum	TRB_sum	AST_sum	PTS_sum	ratio
2	0.33676802	2.8565972	4.1945245	2.79646729	1.94160777	3.04569858	3.85216545	0.4743988	1.84365656	0.17925468
3	-0.4134696	-0.3610726	-0.2999508	-0.9392667	-0.9917394	-0.5813661	-0.6338621	-0.5769912	-0.6680921	0
4	-0.4134696	0.09124939	-0.2999508	-0.5444331	-0.716115	-0.5362015	-0.503813	-0.3479614	-0.3776469	-0.0398986
5	0.87265206	-0.3592787	-0.2999508	-0.7404694	-0.833402	-0.5639951	-0.6081186	-0.4717806	-0.5837191	-0.1633511



A	B	C	D	E	F	G	H	I	J	K
n_height	n_search_num	n_final_win_num	n_G_sum	n_PER_sum	n_WS_sum	n_TRB_sum	n_AST_sum	n PTS_sum	HoF	ratio
0.27414905	2.90562093	3.82930211	2.53635987	1.69334674	2.74532082	3.65328895	0.37638272	1.682569	0	0.17925468
-0.440812	-0.3260716	-0.2972883	-0.9999335	-1.0654768	-0.6096329	-0.6746745	-0.5799232	-0.6999477	0	0
-0.440812	0.12822161	-0.2972883	-0.6261789	-0.8062511	-0.5678567	-0.5492076	-0.3716061	-0.4244462	0	-0.0398986
0.78483549	-0.3242699	-0.2972883	-0.8117494	-0.9165500	-0.5935651	-0.6498228	-0.4842275	-0.6199158	0	-0.1633511

05 Salary Ratio

Regression model

수치 예측 문제

Nba 선수들의 연봉 인상률을
예측하는 것이 목적

전처리 결과

- 총 데이터
13233개, (1203,11)
- 10개**의 독립변수
- 1개의 종속변수
> ratio

모델 학습

- Linear regression
- Decision Tree & Tree Visualization
- Random Forest
- KNN
- Regression summary

모델 평가

- Regression Metrics
- Dummy regressor

05 Salary Ratio

Regression model

결과값 정리

모형	LR	DT	RF	KNN
Training R^2	0.006969	0.9448798	0.8869544	0.11110311
Test R^2	-0.02461	-0.21401	0.33215	-0.364503

05 Salary Ratio

Regression model

결과값 정리

Current function value: 0.620237

Iterations 5

Logit Regression Results

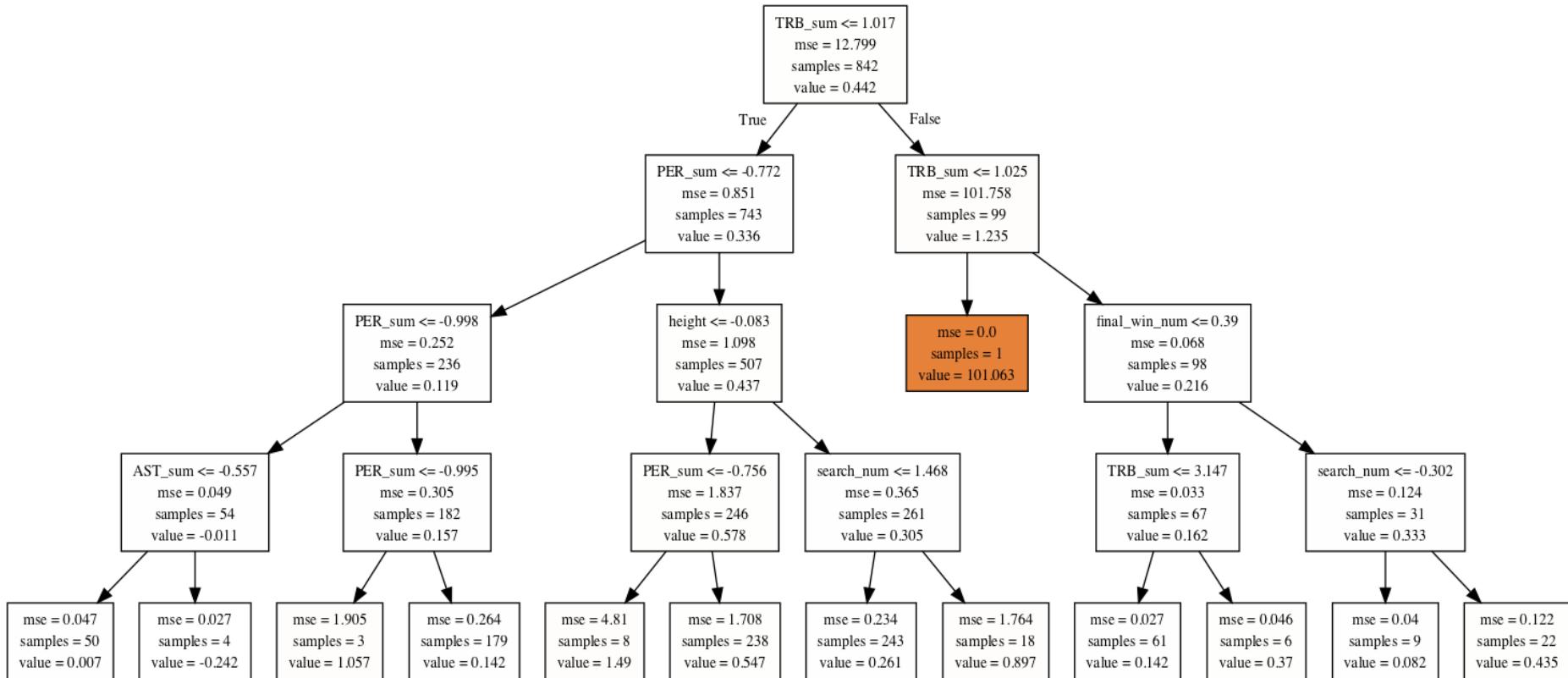
Dep. Variable:	y	No. Observations:	50
Model:	Logit	Df Residuals:	47
Method:	MLE	Df Model:	2
Date:	Sun, 17 Jun 2018	Pseudo R-squ.:	0.1052
Time:	12:42:22	Log-Likelihood:	-31.012
converged:	True	LL-Null:	-34.657
		LLR p-value:	0.02611

	coef	std err	z	P> z	[0.025	0.975]
const	-0.0813	0.308	-0.264	0.792	-0.684	0.522
x1	0.1230	0.065	1.888	0.059	-0.005	0.251
x2	0.1104	0.060	1.827	0.068	-0.008	0.229

05 Salary Ratio

Regression model

Tree Visualization



05 Salary Ratio

Regression model

Regression Metrics

```
[0.3640005 0.20959092 0.4427745 0.23169023 0.4578941 0.40031656  
0.70998996 0.84730011 0.60888876 0.22660138]  
[ 0.18676301 0.34490817 0.15644182 0.15714286 0.02500041 -0.25454545  
0.68442008 10.08636905 0. -0.18233929]  
MAE: 0.7881530380410986  
MSE: 10.489435283587623  
RMSE: 3.2387397678090197  
R^2: -0.02461218032483159
```

Dummy Regressor

```
MAE: 0.7881530380410986  
MSE: 10.489435283587623  
RMSE: 3.2387397678090197
```

05 Salary Ratio 한계

Salary Ratio 의 문제점 분석

✓ 변수 상관관계가 낮음

선수들의 경기 statistics을 비롯한 성적과 개인 정보 및 특징, 화세성은 연봉의 절대값에 영향.

> 경기 성적 등이 낮아도 인상률이 높아지는 결과 발생 : 상관관계 거의 없다.

✓ 부적절한 타겟 값 선정

➤ 유의미하며 적절한 타겟 값 재 선정 - Player 별 Salary

06 새로운 예측값 선정 및 전처리 Salary

전처리 과정

1. 선수의 활동기간 중 연봉액 mean과 기존의 데이터 통합

데이터 통합

salary mean과 Hof (이미 앞에서 전처리한 데이터)를 merge 한다.

#추가적으로 height 등의 데이터를 merge한다

2. 데이터를 표준화한다. > 2가지 방법 사용

데이터 정규화

변수들의 분포 차이가 크므로 정규화 시킨다. Standardscaler

변수들의 분포 차이가 크므로 정규화 시킨다. Minmaxscaling

06 새로운 예측값 선정 및 전처리

Salary

1	n_height	n_search_nun	n_final_win_n	n_G_sum	n_PER_sum	n_WS_sum	n_TRB_sum	n_AST_sum	n PTS sum	HoF	n_salary
2	0.6056338	0.19519453	0.42857143	0.82775411	0.59750173	0.38731884	0.58079128	0.09293939	0.33678068	0	3754136.62
3	0.50704225	3.96E-05	0	0.0042605	0.16360167	0.00905797	0.00126147	0	0.0002084	0	118974
4	0.50704225	0.02747339	0	0.09129641	0.20437196	0.01376812	0.01806193	0.02024548	0.03912783	0	300000
5	0.67605634	0.0001484	0	0.04808278	0.1870229	0.01086957	0.00458716	0.00930027	0.01151431	0	267189
6	0.42253521	4.84E-06	0	0.00304321	0.17748092	0.00978261	0.00017202	0.00050614	0.00031261	0	30000
7	0.50704225	0.0003064	0	0.58186245	0.5019084	0.17246377	0.17425459	0.15576363	0.17287102	0	2566467.05
8	0.35211268	0.03819346	0	0.01095557	0.17470507	0.00978261	0.00080275	0.00151841	0.00041681	0	155614
9	0.50704225	4.12E-05	0	0.00486914	0.18459403	0.00978261	0.00045872	0.0001898	0.00062521	0	102424
10	0.64788732	0.0735937	0	0.50943396	0.55013879	0.12463768	0.18199541	0.03321523	0.12030114	0	2016724.79
11	0.6056338	3.68E-05	0	0.08642727	0.22362942	0.02934783	0.02866972	0.01012274	0.03360513	0	0

Min-Max 후 CSV

A	B	C	D	E	F	G	H	I	J	K	
1	n_height	n_search_nun	n_final_win_n	n_G_sum	n_PER_sum	n_WS_sum	n_TRB_sum	n_AST_sum	n PTS sum	HoF	n_salary
2	0.33676802	2.8565972	4.1945245	2.79646729	1.94160777	3.04569858	3.85216545	0.4743988	1.84365656	0	3754136.62
3	-0.4134696	-0.3610726	-0.2999508	-0.9392667	-0.9917394	-0.5813661	-0.6338621	-0.5769912	-0.6680921	0	118974
4	-0.4134696	0.09124939	-0.2999508	-0.5444331	-0.716115	-0.5362015	-0.503813	-0.3479614	-0.3776469	0	300000
5	0.87265206	-0.3592787	-0.2999508	-0.7404694	-0.833402	-0.5639951	-0.6081186	-0.4717806	-0.5837191	0	267189
6	-1.0565305	-0.3616457	-0.2999508	-0.9447889	-0.8979098	-0.5744177	-0.6422953	-0.5712654	-0.6673145	0	300000
7	-0.4134696	-0.3566736	-0.2999508	1.68099313	1.29535655	0.98549805	0.70524401	1.18510694	0.62044274	0	2566467.05

standardization 후 CSV

06 Salary

Regression model

수치 예측 문제

Nba 선수들의 **연봉**을
예측하는 것이 목적

전처리 결과

Min-Max/ Standard 모두

- 총 데이터
20658개, (1878,11)
- 10개**의 독립변수
- 1개의 종속변수
> **n_salary**

모델 학습

- Linear regression**
- Decision Tree & Tree Visualization**
- Random Forest**
- KNN**
- Regression summary**

모델 평가

- Regression Metrics**
- Dummy regressor**

06 Salary

Regression model

결과값 정리

Min-Max & Standardization

연1	LR	DT	RF	KNN
Training R	0.314387	0.362463	0.8869544	0.348615
Test R^	0.273071	0.29869	0.33215	0.1891409

```
intercept: 856025.1754627451
coefficient: [-3039156.47736018  126620.42258313 -85536.19609396 -79502.26021394
 -347340.46371652  928695.09460198  142343.50078734  27723.49032266
 14787.37340024  94652.03439414]
height -3039156.47736018
search_num 126620.4225831256
final_win_num -85536.19609396029
G_sum -79502.2602139373
PER_sum -347340.4637165223
WS_sum 928695.0946019753
TRB_sum 142343.50078733749
AST_sum 27723.490322662234
PTS_sum 14787.373400237819
HoF 94652.03439413734
```

06 Salary

Regression model

결과값 정리

Current function value: 0.620237

Iterations 5

Logit Regression Results

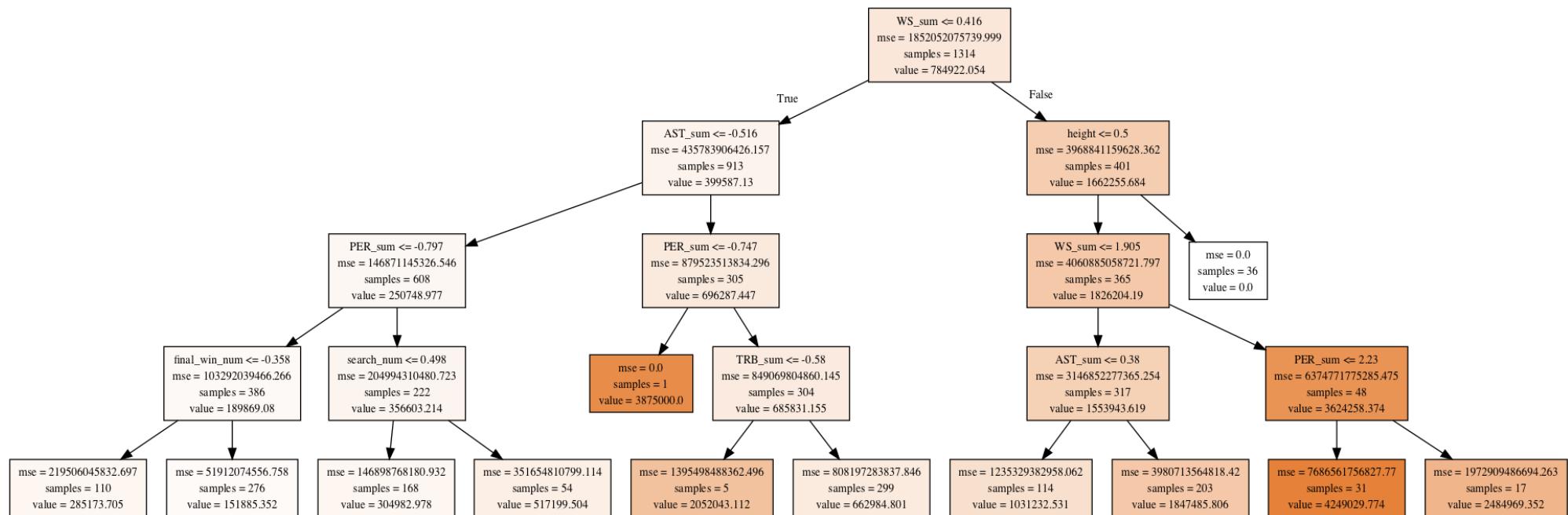
Dep. Variable:	y	No. Observations:	50
Model:	Logit	Df Residuals:	47
Method:	MLE	Df Model:	2
Date:	Sun, 17 Jun 2018	Pseudo R-squ.:	0.1052
Time:	11:43:36	Log-Likelihood:	-31.012
converged:	True	LL-Null:	-34.657
		LLR p-value:	0.02611

	coef	std err	z	P> z	[0.025	0.975]
const	-0.0813	0.308	-0.264	0.792	-0.684	0.522
x1	0.1230	0.065	1.888	0.059	-0.005	0.251
x2	0.1104	0.060	1.827	0.068	-0.008	0.229

06 Salary

Regression model

Tree Visualization



06 Salary

Regression 모델 평가 Report

Feature Selection – 모델 기반 선택 mean/ median

전체 변수 사용: 0.2772975169256249
선택 변수 사용: 0.0776574865880475

전체 변수 사용: 0.2772975169256249
선택 변수 사용: 0.11685538775685289

Feature Selection – 반복적 선택 RFE / feature select가 1,3,5 인 경우

total_score: 0.2772975169256249
rfe_score: -0.4467036338503568

total_score: 0.2772975169256249
rfe_score: 0.12972437941626613

total_score: 0.2772975169256249
rfe_score: 0.11685538775685289

06 Salary

Regression 모델 평가 Report

Regression metrics

```
[ 150688.1340641  949488.49859119  851260.94090124  870585.06349535
 223093.76163006 1344561.61614021 -467567.86673939  740366.84341855
1087923.12100764 689462.28266904]
[      0.          0.        2097025.       525000.       574566.       1299381.
      0.          0.       370887.25  1469908.571]
MAE: 615825.2241678645
MSE: 900653068174.2266
RMSE: 949027.432782755
R^2: 0.2730717935250341
```

dummy regressor

```
MAE: 615825.2241678645
MSE: 900653068174.2266
RMSE: 949027.432782755
```

06 결과값 향상 전개 과정

예측 결과값 정리

열1	LR	DT	RF	KNN
Training R	0.006969	0.9448798	0.8869544	0.11110311
Test R^	-0.02461	-0.21401	0.33215	-0.364503

Salary Ratio 연봉 인상률

열1	LR	DT	RF	KNN
Training R	0.314387	0.362463	0.8869544	0.348618
Test R^	0.273071	0.29869	0.33215	0.189140

Salary

07 결론 및 시사점

시사점 및 프로젝트 후기

Hall of Fame

- ✓ Over sampling 통해 유의미한 예측값 산출
- ✓ 각 변수와의 상관관계

Salary

- ✓ 결과값은 낮으나 regression 임을 감안
- ✓ 각 변수와의 상관관계

[결과]

[시사점]

- I. NBA 선수들의 미래 성공 가능성 지표가 된다.
- II. 광고회사에서 미리 Hall Of Fame에 들어 갈 선수의 명단을 활용해 해당 선수 선점 가능하다. > 광고 비용 절감할 수 있다.
- III. 새로 유입된 농구 팬들에게 선수나 팀 선택 기준이 된다.

- I. 구단의 입장에서 트레이드나 재계약 시 유용하며, 자원을 구분할 수 있는 지표가 된다.
- II. FA 시장에 나온 선수 계약 시 활용 가능하다

07 결론 및 시사점

시사점 및 프로젝트 후기



생소했던 데이터 마이닝과 파이썬을 쉽게 그리고 자세
하게 배울 수 있어 정말 좋았습니다!
이제 조금은 파이썬을 알 것 같아요:)



틈틈히 있는 과제와 마지막 프로젝트가 살짝 ... 버겁긴
했으나 그래도 한 학기만에 파이썬&데이터마이닝과 좀
친해진 기분입니다. 재밌었고, 드디어 스파이더 종료

파이썬과 데이터 마이닝 기초를 배우자마자 데
이터를 모델링해보는 스파르타식 수업이었으나
얻어가는 게 가장 많은 수업이었습니다. 최공.



나는 4차 산업혁명보다
고대 그리스가 더 어울리나보다



감사합니다 :D