NBA 연봉 예측 분석 프로젝트

이 프로젝트는 NBA 선수들의 경기 통계(득점, 어시스트, 리바운드, 슈팅 성공률 등)를 활용하여 연봉을 예측하는 것을 목표로 합니다. 회귀 모델과 고급 특징 엔지니어링 기법을 활용하여 선수 연봉에 영향을 미치는 주요 요인을 분석하고 예측합니다.

1. 데이터 수집

이 프로젝트에서는 Selenium 과 BeautifulSoup 을 사용하여 웹 크롤링을 통해 데이터를 수집하였습니다. 수집된 데이터는 nba_stats_yyyy-yy.csv, nba_salary_yyyy-yy.csv 형식으로 저장됩니다.

사용된 기술

- Selenium: 동적 페이지 렌더링을 처리하여 데이터를 크롤링
- BeautifulSoup: HTML 문서 구조를 파싱하여 원하는 데이터를 추출

NBA 선수 개인 지표

https://www.nba.com/stats/players/traditional

488 Rows • Page 1 of 10

	PLAYER	TEAM	AGE	GP	W	L	MIN	PTS	FGM	FGA	FG%	3PM	3PA	3P%	FTM	FTA	FT%	OREB	DREB	REB	AST	TOV	STL	BLK	PF	FP	DD2	TD3	+/-
1	Giannis Antetokounmpo	MIL	29	16	8	8	35.0	32.4	13.1	21.5	60.8	0.2	0.9	21.4	6.1	10.1	60.2	2.1	9.8	11.9	6.4	3.3	0.6	1.4	2.9	59.0	14.0	2.0	-0.1
2	LaMelo Ball	CHA	23	18	6	12	34.1	31.1	10.7	24.9	43.0	4.7	13.1	35.6	4.9	5.8	84.8	0.9	4.4	5.4	6.9	4.5	1.1	0.2	4.1	47.4	4.0	0.0	-0.9
3	Nikola Jokić	DEN	29	14	9	5	37.3	29.7	11.1	19.5	56.8	2.2	4.1	53.4	5.4	6.4	83.3	4.2	8.9	13.1	10.6	3.6	1.5	0.9	1.8	64.9	12.0	7.0	10.1
4	Shai Gilgeous-Alexander	OKC	26	18	14	4	34.3	29.5	10.3	20.3	50.5	2.0	5.8	34.3	6.9	8.1	85.6	0.8	4.5	5.3	6.4	2.8	1.7	1.1	1.8	51.0	2.0	0.0	10.6
5	Anthony Davis	LAL	31	17	11	6	35.2	29.2	10.4	18.8	55.5	0.8	2.1	36.1	7.6	9.7	78.2	2.7	8.8	11.5	3.2	2.1	1.3	1.9	2.0	55.4	11.0	0.0	-1.2
6	Paolo Banchero	ORL	22	5	3	2	36.4	29.0	9.6	19.4	49.5	2.2	6.4	34.4	7.6	11.8	64.4	2.4	6.4	8.8	5.6	2.2	0.6	0.8	2.6	50.0	2.0	0.0	5.8

- requests 와 BeautifulSoup 을 사용하여 HTML 페이지를 가져옵니다.

- 테이블 데이터를 파싱하고 pandas 를 사용하여 데이터프레임 형식으로 저장합니다.

- 페이지네이션 처리로 모든 급여 데이터를 수집합니다

아래는 데이터를 수집하기 위해 사용된 주요 코드입니다:

NBA 선수 연봉

https://www.nba.com/stats/players/traditional

2024-2025 Player Salaries			
RK	NAME	TEAM	SALARY
1	Stephen Curry, PG	Golden State Warriors	$55,761,216
2	Joel Embiid, C	Philadelphia 76ers	$51,415,938
3	Nikola Jokic, C	Denver Nuggets	$51,415,938
4	Kevin Durant, PF	Phoenix Suns	$51,179,021
5	Bradley Beal, SG	Phoenix Suns	$50,203,930
6	Kawhi Leonard, SF	LA Clippers	$49,350,000
7	Devin Booker, SG	Phoenix Suns	$49,205,800
8	Paul George, F	Philadelphia 76ers	$49,205,800
9	Karl-Anthony Towns, C	New York Knicks	$49,205,800
10	Jaylen Brown, SG	Boston Celtics	$49,205,800

- Selenium 을 사용하여 동적 웹 페이지에서 데이터를 자동으로 로드합니다.
- 드롭다운 메뉴를 조작하여 시즌 및 필터 조건을 설정.
- 데이터를 추출한 뒤, pandas 를 사용하여 데이터프레임 형식으로 변환합니다.

2. 데이터 전처리 및 변환

전처리 전 연봉 데이터

	RK	NAME	TEAM	SALARY
0	1	Stephen Curry, PG	Golden State Warriors	$55,761,216
1	2	Joel Embiid, C	Philadelphia 76ers	$51,415,938
2	3	Nikola Jokic, C	Denver Nuggets	$51,415,938
3	4	Kevin Durant, PF	Phoenix Suns	$51,179,021
4	5	Bradley Beal, SG	Phoenix Suns	$50,203,930
...	...	...	...	...
460	461	Tyler Smith, F	Milwaukee Bucks	$1,157,153
461	462	Bronny James, G	Los Angeles Lakers	$1,157,153
462	463	Cam Christie, G	LA Clippers	$1,157,153
463	464	Antonio Reeves, G	New Orleans Pelicans	$1,157,153
464	465	Pelle Larsson, G	Miami Heat	$1,157,143

- 선수 이름에 포함되어 있는 포지션 제거

```
# 연봉 데이터의 이름에서 포지션 제거
salary_data['NAME'] = salary_data['NAME'].str.split(',').str[0].str.strip()
```

- 연봉을 숫자 형식으로 전처리

```
 # 연봉 데이터에서 $ 제거 및 숫자형으로 변환
 salary_data['SALARY'] = salary_data['SALARY'].replace(r'[$,]', '', regex=True).astype(int)
```

- 샐러리 캡 기준으로 조정

```
# 각 선수의 연봉이 시즌별 샐러리 캡에서 차지하는 비율 계산 SALARY_CAP_RATIO
# SALARY_CAP_RATIO과 현재 년도의 샐러리 캡을 이용해 예상 연봉 계산 CURRENT_SALARY_PROJECTION
# 'CURRENT_SALARY_PROJECTION'는 선수 연봉의 상대적 가치를 평가하는 데 사용. (종속변수)
currunt_salary_cap = salary_cap.get(salary_year)
salary_data['ADJUSTED_SALARY_FOR_CAP'] = (salary_data['SALARY'] / currunt_salary_cap * base_salary_cap).astype(int)
```

전처리 전 개인 지표 데이터

	Unnamed: 0	Player	Team	Age	GP	W	L	Min	PTS	FGM	...	REB RANK	AST RANK	TOV RANK	STL RANK	BLK RANK	PF RANK	FP RANK	DD2 RANK	TD3 RANK	+/- RANK
0	1	Joel Embiid	PHI	30	39	31	8	33.6	34.7	11.5	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
1	2	Luka Dončić	DAL	25	70	46	24	37.5	33.9	11.5	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2	3	Giannis Antetokounmpo	MIL	29	73	45	28	35.2	30.4	11.5	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
3	4	Shai Gilgeous-Alexander	OKC	25	75	55	20	34.0	30.1	10.6	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
4	5	Jalen Brunson	NYK	27	77	49	28	35.4	28.7	10.3	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
567	563	Justin Jackson	MIN	29	2	2	0	0.4	0.0	0.0	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
568	563	Kaiser Gates	NOP	27	1	0	1	7.4	0.0	0.0	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
569	563	Malcolm Cazalon	DET	22	1	0	1	2.6	0.0	0.0	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
570	563	Ron Harper Jr.	TOR	24	1	1	0	3.7	0.0	0.0	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
571	563	Ryan Arcidiacono	NYK	30	20	15	5	2.3	0.0	0.0	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

- 이름 악센트 제거

```
# 악센트 제거 함수
def remove_accents(input_str):
    return ''.join(
        char for char in unicodedata.normalize('NFD', input_str)
        if unicodedata.category(char) != 'Mn'  # Nonspacing Mark 제외
    )
# 악센트가 포함되어 있는 선수 악센트 제거
# 연봉 데이터에는 악센트가 이미 제거 되어있어 개인 지표 데이터에만 실행
stats_data['Player'] = stats_data['Player'].apply(remove_accents)
```

- 선수 이름을 기준으로 개인 기록 데이터와 연봉 데이터를 병합

```
# 이름을 기준으로 병합
merged_data = pd.merge(
    stats_data, salary_data,
    how='inner',
    left_on=['Player'],
    right_on=['NAME']
)
```

- 시즌 열을 추가후 리스트에 결합된 데이터 추가

```
# 시즌 정보 추가
merged_data['SEASON'] = f'{salary_year}-{salary_year+1}'

# 병합된 데이터 추가
all_data.append(merged_data)
```

- 모든 시즌 데이터를 하나의 데이터 프레임으로 결합 후 결측치 제거

```
# 모든 시즌 데이터를 하나의 데이터프레임으로 결합
final_data = pd.concat(all_data, ignore_index=True)
final_data.dropna(axis=1, inplace=True)
```

- 불필요한 독립 변수 제거

```
final_data.drop(columns=['Unnamed: 0', 'Team', 'W', 'L', 'FGM', 'FGA', 'PF', 'FP', 'OREB',
                         'DREB', '3PM', '3PA', 'FTA', 'FTM', 'TEAM', 'NAME', 'RK'], axis=1, inplace=True)
```

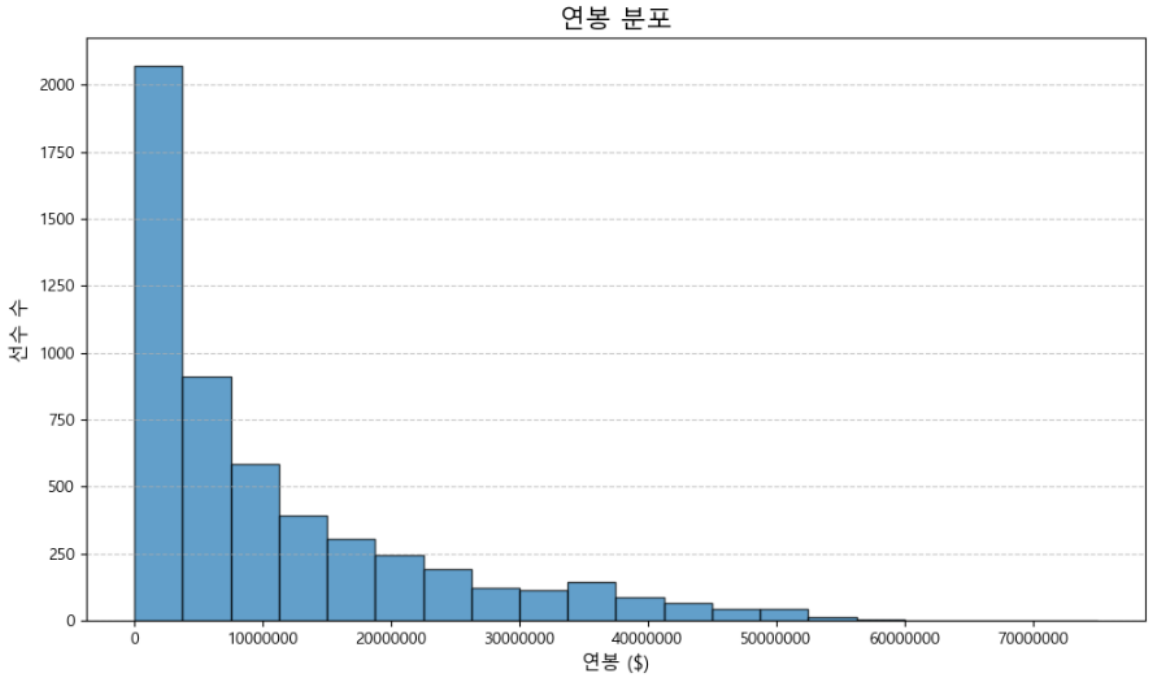
- 최종 데이터

	Player	Age	GP	Min	PTS	FG%	3P%	FT%	REB	AST	TOV	STL	BLK	DD2	TD3	+/-	SALARY	ADJUSTED_SALARY_FOR_CAP	SEASON
0	Kevin Durant	23	66	38.6	28.0	49.6	38.7	86.0	8.0	3.5	3.8	1.3	1.2	18.0	0.0	5.6	17832626	43933318	2012-2013
1	Kobe Bryant	33	58	38.5	27.9	43.0	30.3	84.5	5.4	4.6	3.5	1.2	0.3	3.0	0.0	2.4	30453805	75027463	2012-2013
2	LeBron James	27	62	37.5	27.1	53.1	36.2	77.1	7.9	6.2	3.4	1.9	0.8	23.0	0.0	7.6	19067500	46975613	2012-2013
3	Kevin Love	23	55	39.0	26.0	44.8	37.2	82.4	13.3	2.0	2.3	0.9	0.5	48.0	0.0	0.5	13668750	33674992	2012-2013
4	Russell Westbrook	23	66	35.3	23.6	45.7	31.6	82.3	4.6	5.5	3.6	1.7	0.3	7.0	0.0	5.6	14693906	36200616	2012-2013
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
5332	DaQuan Jeffries	26	17	2.7	0.8	35.3	20.0	0.0	0.3	0.3	0.2	0.0	0.1	0.0	0.0	-0.6	2425204	2425204	2024-2025
5333	Charlie Brown Jr.	27	8	4.7	0.8	20.0	28.6	0.0	0.3	0.0	0.3	0.0	0.3	0.0	0.0	-0.8	2237692	2237692	2024-2025
5334	Wendell Moore Jr.	22	25	3.0	0.7	50.0	0.0	0.0	0.5	0.2	0.2	0.2	0.0	0.0	0.0	-0.2	2537040	2537040	2024-2025
5335	E.J. Liddell	23	8	2.9	0.5	16.7	0.0	100.0	0.6	0.1	0.3	0.3	0.3	0.0	0.0	-0.5	2120693	2120693	2024-2025
5336	Maxwell Lewis	21	34	3.0	0.3	19.0	11.1	66.7	0.1	0.2	0.3	0.1	0.0	0.0	0.0	-2.1	1891857	1891857	2024-2025

3. 시각화 및 분석 결과

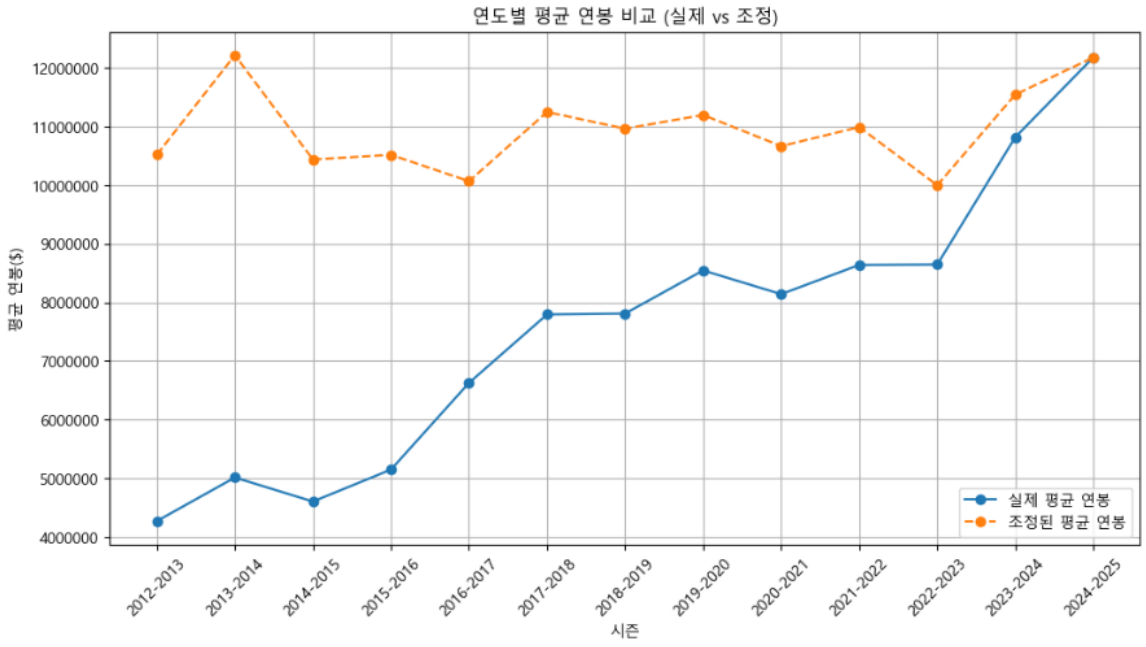
- 연봉 분포

대부분의 선수 연봉이 특정 범위에 집중되어 있음을 확인 할 수 있다. 연봉의 분포는 비대칭적으로 나타나며, 일부 상위 선수들이 매우 높은 연봉을 받고 있는 것을 보여준다.



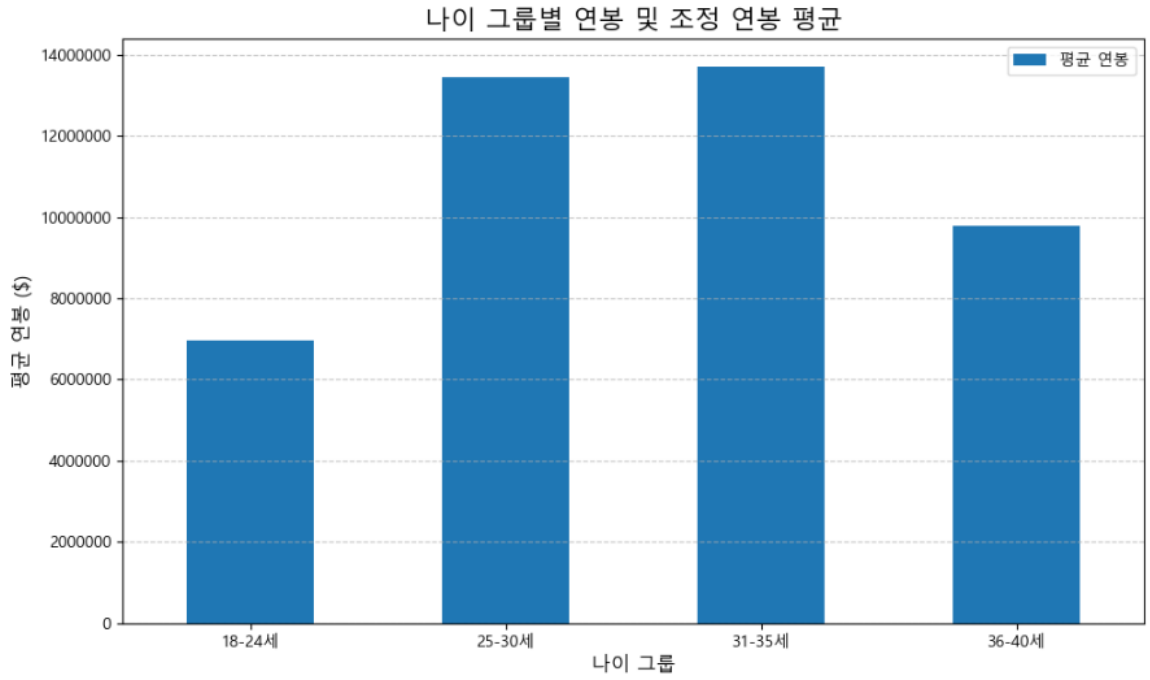
- 시즌별 연봉

라인 그래프를 통해 시즌별로 연봉 상승을 확일 할 수 있다. 또한 조정 전 연봉은 편차가 심한 것을 볼 수 있는데 이를 샐러리 캡 계산을 통해 편차를 줄인 것을 확일 할 수 있다.

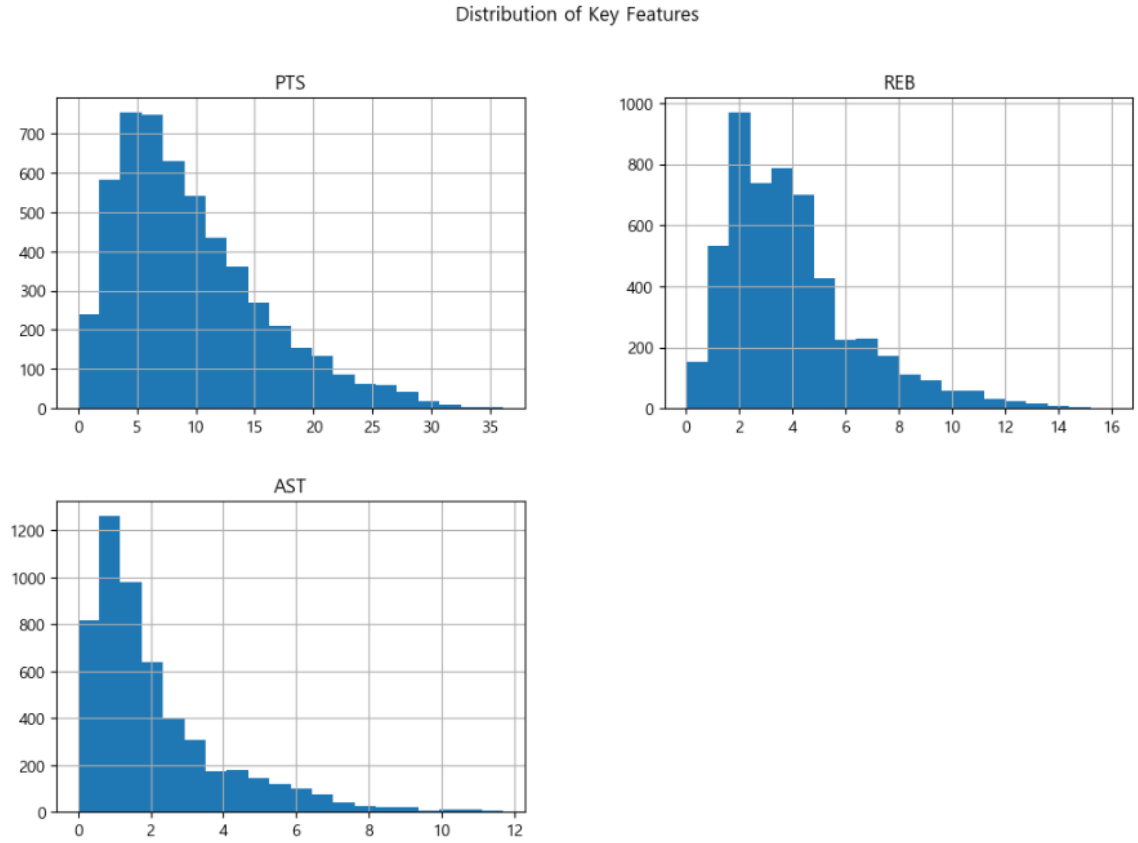


- 나이대별 평균 연봉

- 25-30 세 그룹이 가장 높은 평균 연봉을 기록하며, 이는 일반적으로 선수들이 전성기(peak performance)를 맞는 시기임을 나타냄.
- 18-24 세 그룹은 신인 계약 또는 경험 부족으로 인해 상대적으로 낮은 연봉을 기록.
- 31-35 세 그룹의 연봉은 여전히 경쟁력을 유지하지만, 연령 증가에 따른 경기력 저하로 25-30 세 그룹보다 낮음.

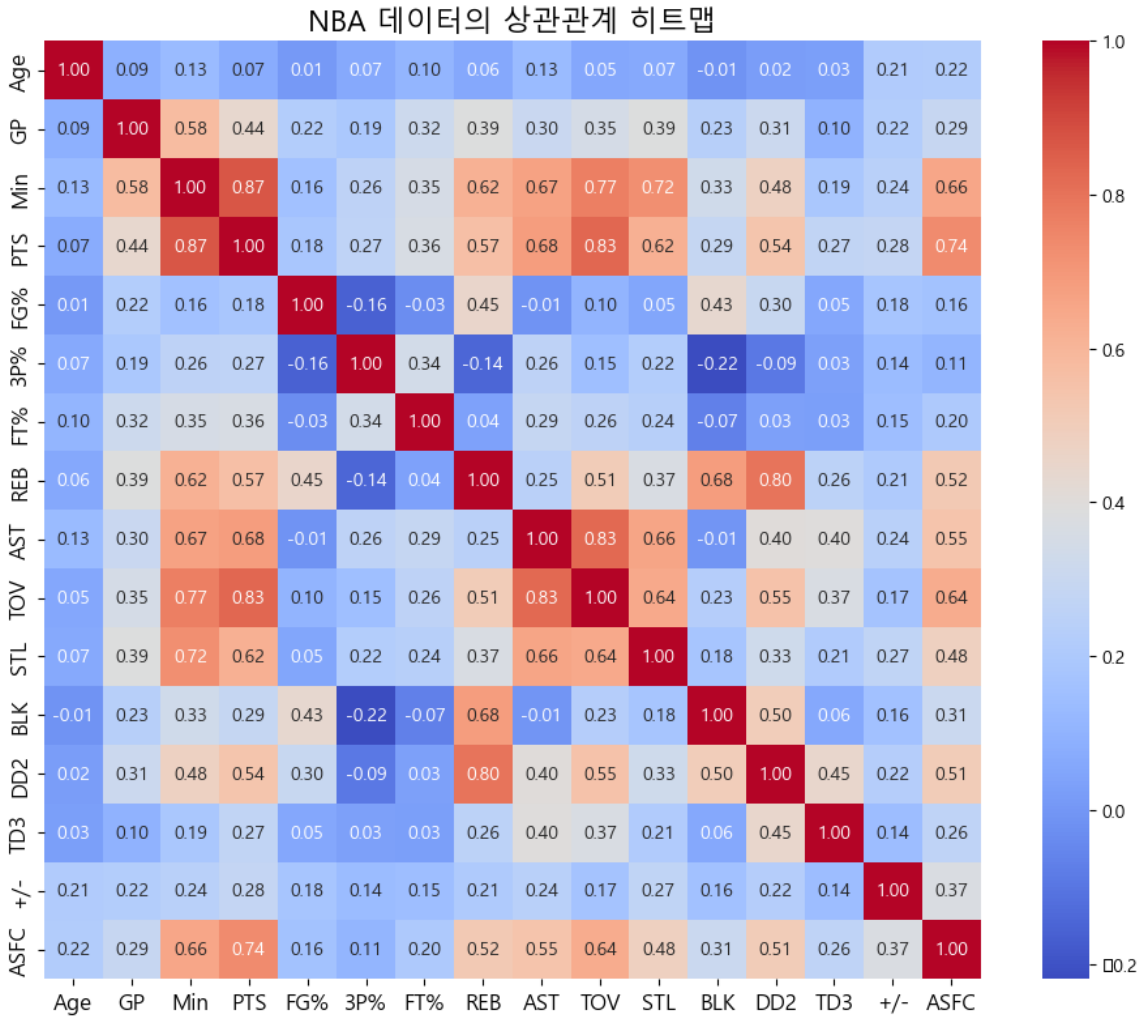


- 주요 변수 분포



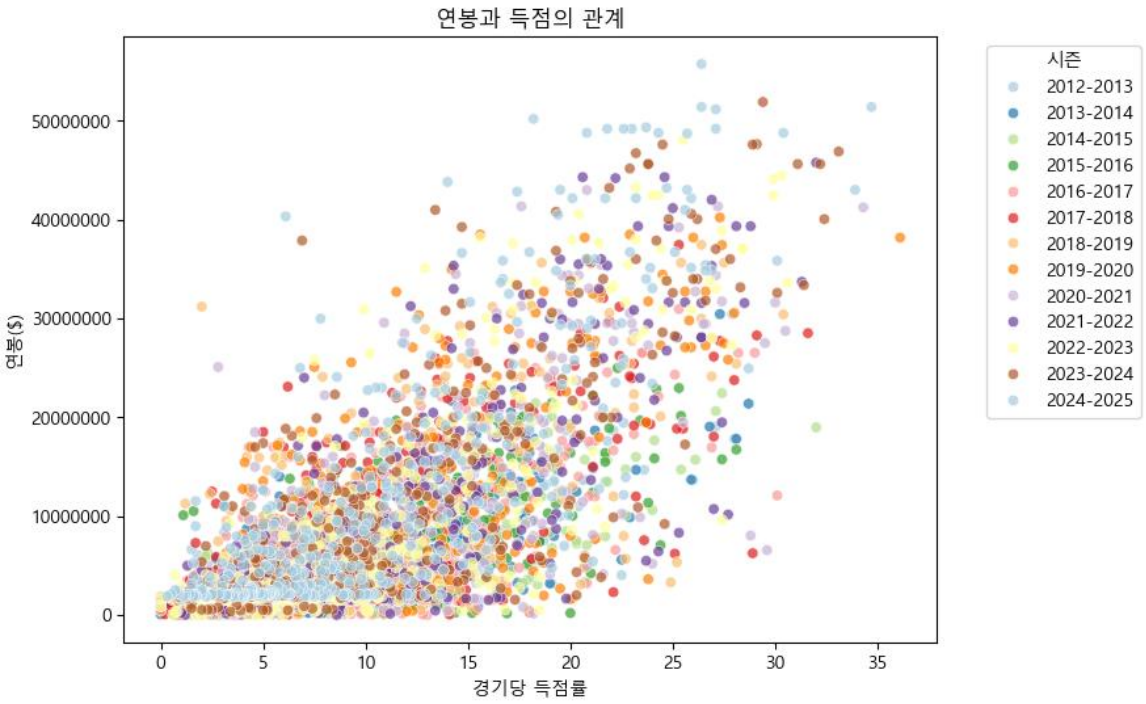
- 연봉과 지표 간의 상관관계 분석

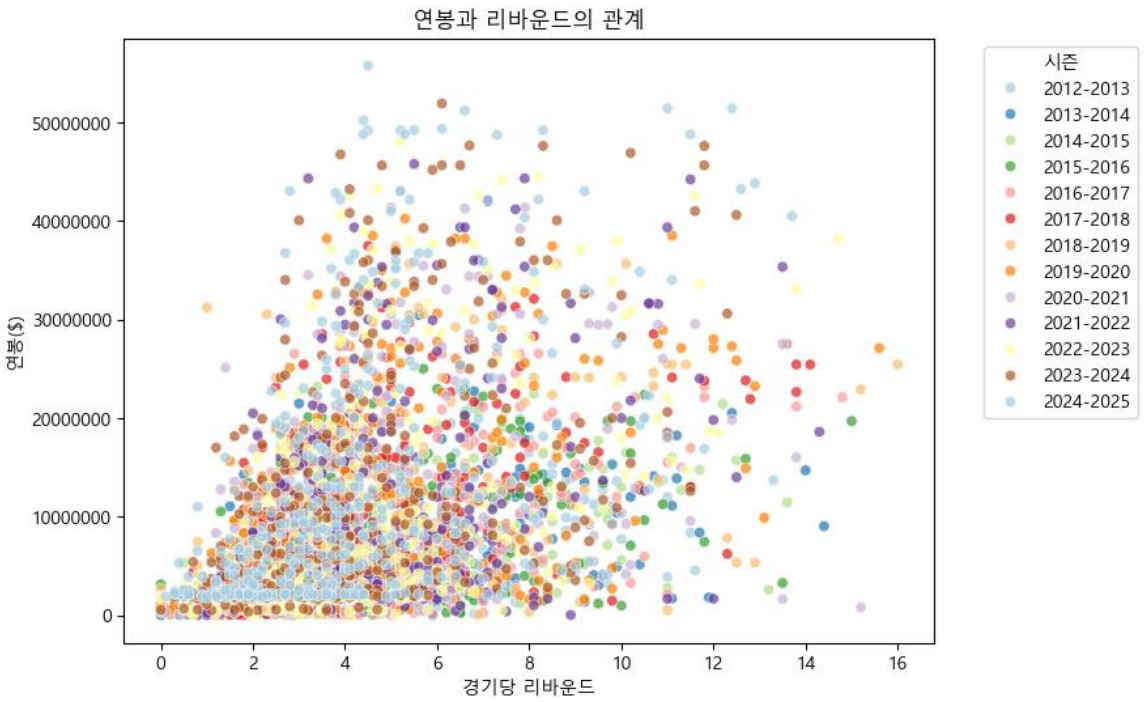
득점(PTS)은 연봉과 가장 강한 상관관계를 가지며, 이는 득점력이 높은 선수가 더 높은 연봉을 받는 경향이 있다는 것을 나타낸다.

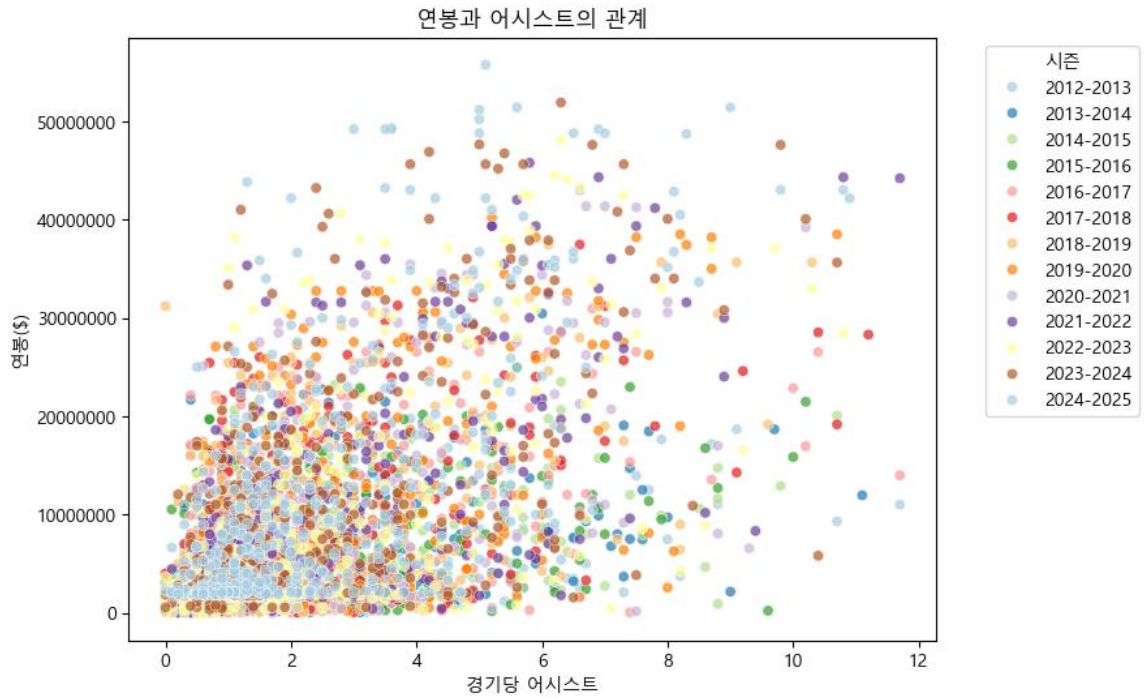


- 득점, 리바운드, 어시스트와 연봉의 관계.

- 산점도를 통해 경기당 득점이 높은 선수일수록 연봉이 높아지는 경향을 확인할 수 있다.
- 리바운드와 어시스트는 포지션에 따른 차이가 있을 수 있다. (예: 키가큰 빅맨 포지션은 연봉이 적어도 리바운드 횟수가 높을 수 있음.)




연봉과 리바운드의 관계
연봉($)
경기당 리바운드
시즌
2012-2013
2013-2014
2014-2015
2015-2016
2016-2017
2017-2018
2018-2019
2019-2020
2020-2021
2021-2022
2022-2023
2023-2024
2024-2025


연봉과 어시스트의 관계
연봉($)
경기당 어시스트
시즌
2012-2013
2013-2014
2014-2015
2015-2016
2016-2017
2017-2018
2018-2019
2019-2020
2020-2021
2021-2022
2022-2023
2023-2024
2024-2025

4. 모델링 및 성능 평가

다양한 회귀 모델을 사용하여 연봉 예측 모델을 학습했습니다:

독립 변수 데이터

	Age	GP	Min	PTS	FG%	3P%	FT%	REB	AST	TOV	STL	BLK	DD2	TD3	+/-
0	23	66	38.6	28.0	49.6	38.7	86.0	8.0	3.5	3.8	1.3	1.2	18.0	0.0	5.6
1	33	58	38.5	27.9	43.0	30.3	84.5	5.4	4.6	3.5	1.2	0.3	3.0	0.0	2.4
2	27	62	37.5	27.1	53.1	36.2	77.1	7.9	6.2	3.4	1.9	0.8	23.0	0.0	7.6
3	23	55	39.0	26.0	44.8	37.2	82.4	13.3	2.0	2.3	0.9	0.5	48.0	0.0	0.5
4	23	66	35.3	23.6	45.7	31.6	82.3	4.6	5.5	3.6	1.7	0.3	7.0	0.0	5.6
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
5332	26	17	2.7	0.8	35.3	20.0	0.0	0.3	0.3	0.2	0.0	0.1	0.0	0.0	-0.6
5333	27	8	4.7	0.8	20.0	28.6	0.0	0.3	0.0	0.3	0.0	0.3	0.0	0.0	-0.8
5334	22	25	3.0	0.7	50.0	0.0	0.0	0.5	0.2	0.2	0.2	0.0	0.0	0.0	-0.2
5335	23	8	2.9	0.5	16.7	0.0	100.0	0.6	0.1	0.3	0.3	0.3	0.0	0.0	-0.5
5336	21	34	3.0	0.3	19.0	11.1	66.7	0.1	0.2	0.3	0.1	0.0	0.0	0.0	-2.1

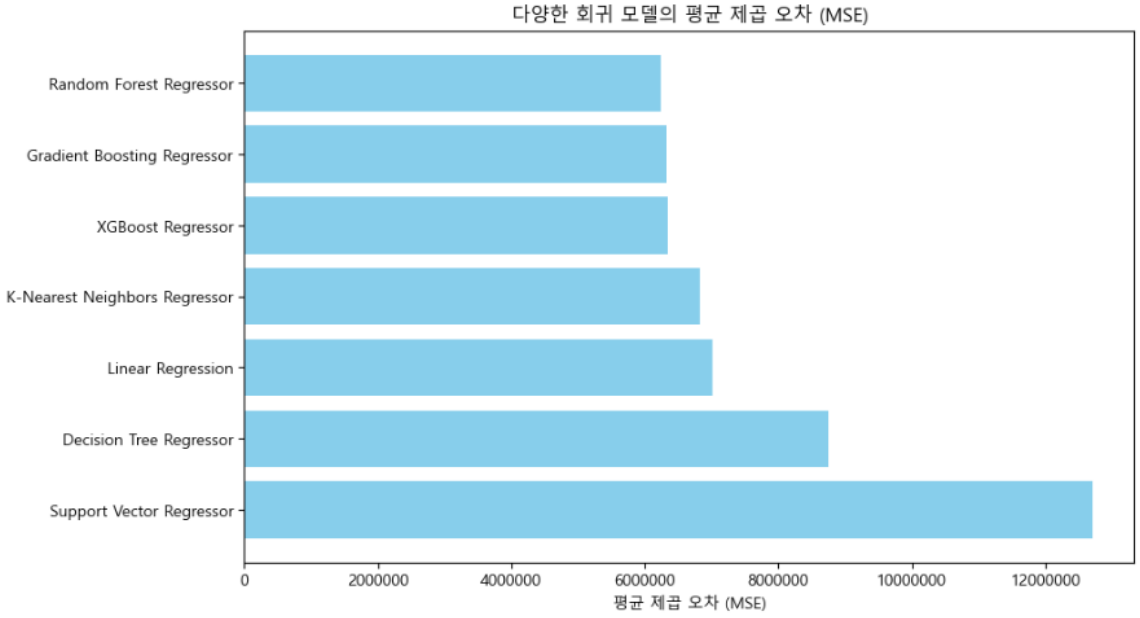
종속 변수 데이터

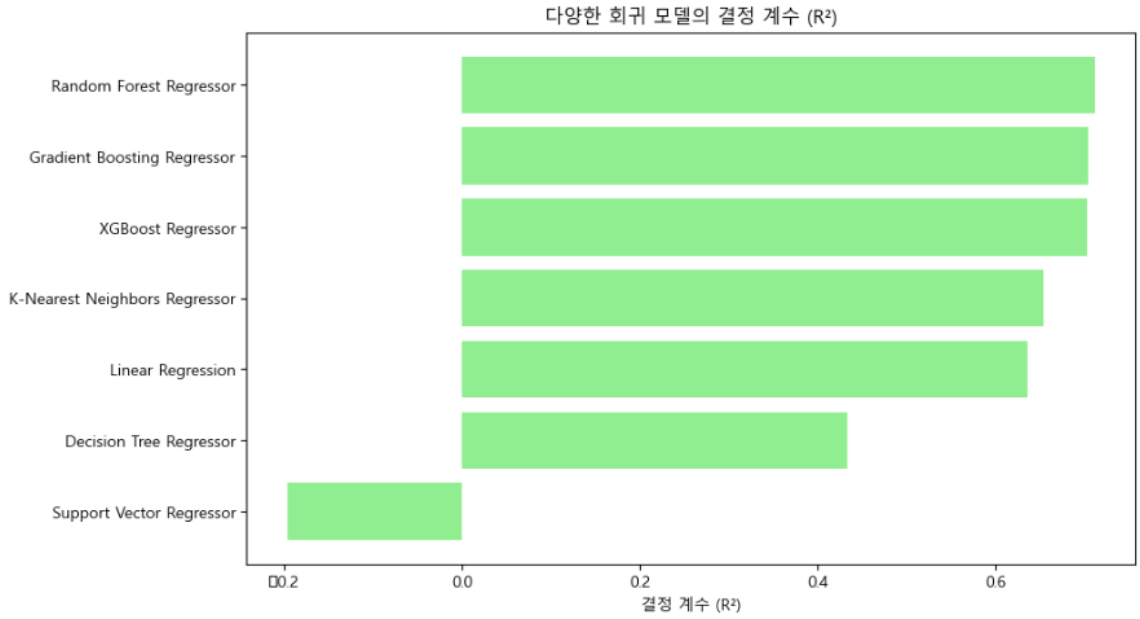
	ADJUSTED_SALARY_FOR_CAP
0	43933318
1	75027463
2	46975613
3	33674992
4	36200616
...	...
5332	2425204
5333	2237692
5334	2537040
5335	2120693
5336	1891857

5337 rows × 1 columns

- 사용 모델 선형 회귀, 랜덤 포레스트, Gradient Boosting 등
- 평균 제곱 오차(MSE)와 결정 계수(R^2)를 사용하여 모델 성능 평가

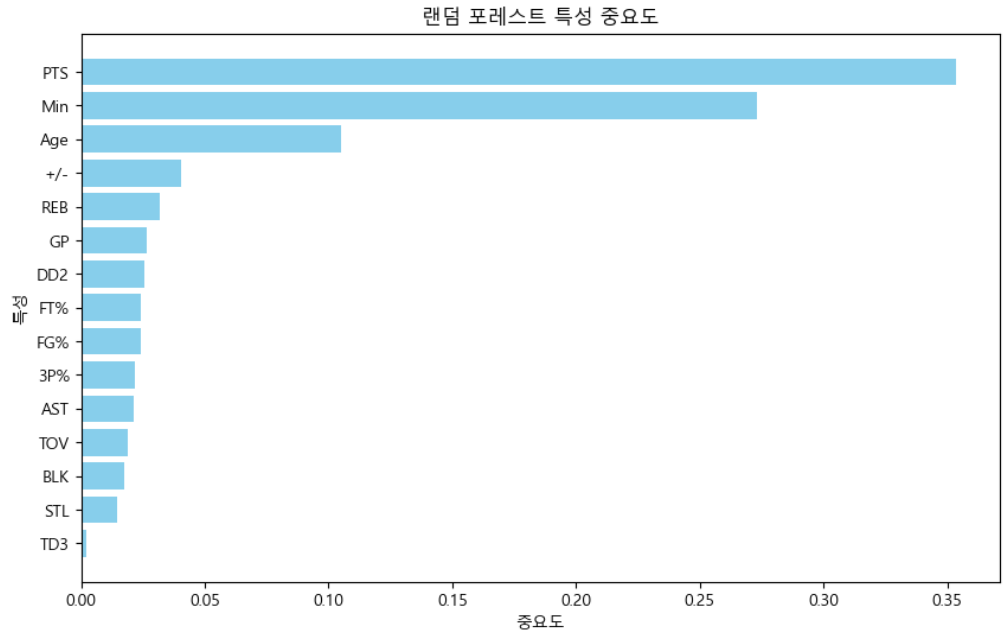
- 랜덤 포레스트 모델이 가장 우수한 성능을 보인다.



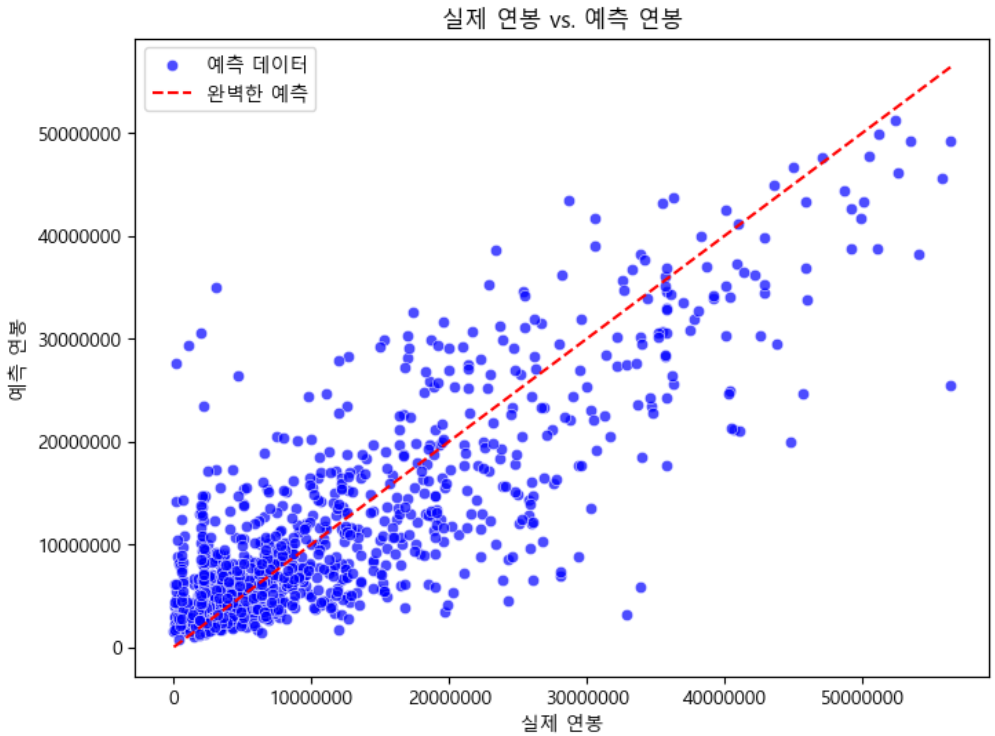


- 랜덤 포레스트를 이용한 특성 중요도

득점률이 연봉에 가장 큰 영향을 미치는 변수로 나타난다.



- 실제 연봉과 예측 연봉 비교



5. 결론

득점은 연봉 결정에 있어 가장 중요한 변수로 나타났으며, 이는 공격력이 NBA 에서 선수 가치를 평가하는 주요 기준임을 의미한다.

실제 연봉과 예측된 연봉을 비교하여 저평가 된 선수와 고평가 된 선수를 알 수 있었다. 이러한 정보를 통해 팀은 더 효율적으로 샐러리 캡을 활용하거나, 재계약 시 전략적 결정을 내릴 수 있다.

외부 요인(예: 마케팅 가치, 팬 영향력, 부상 등) 을 고려한 데이터가 있다면 좀 더 정확한 연봉을 예측할 수 있을 것이다.

6. GitHub 링크

프로젝트 전체 코드는 GitHub 에서 확인할 수 있습니다:

https://github.com/hwangjihong/nba_salary_prediction