

데이터 영향이 직군 높은 만족도에 직군 높은 만족도에 요인 분석

AI 02 김진우



Contents

- 1- 선정 이유 및 데이터 개요**
- 2- 데이터 최적화(전처리)**
- 3- 데이터 시각화 및 요인 분석**
- 4- 머신러닝을 활용한 분석**
- 5- 프로젝트 인사이트 및 회고**



1. 선정 이유 및 데이터 개요



- 요즘 AI기술의 발달로 데이터 사이언티스트, 분석가, 엔지니어, 비지니스 분석가 등의 직군이 주목받고 있음
- 현업에서 근무하는 사람들은 본인들 직업에 만족하는지? 만족한다면, 어떤 부분이 만족도에 영향이 높은지를 분석하는 프로젝트

<분석내용>

- 연봉, 기업 매출, 직원 규모, 업종, 사업 부분, 본/지사 위치, 회사 유형에 따른 만족도 분석 및 시각화
- 머신 러닝 모델을 구축하여 어떤 특성이 만족도에 긍/부정 영향이 있는지 분석



1. 선정 이유 및 데이터 개요

데이터 : 미국 데이터 과학자, 분석가, 엔지니어, 비지니스 분석가 구인광고

데이터 갯수 (12,172ea)

- Data Analyst : 2253ea
- Data Scientist : 3909ea
- Data Engineer : 2528ea
- Business Analyst : 4902ea

특성정보 (15 features)

- Job Title : 직업 타이틀
- Salary Estimate : 연봉
- Job Description : 직업 설명
- Rating : 평가 등급
- Company Name : 회사명
- Location : 직장 위치
- Headquarters : 본사 위치
- Size : 직원 규모
- Founded : 설립시기
- Type of ownership : 회사유형(사기업, 공기업, 공공기관 등)
- Industry : 업종
- Sector : 사업 부문
- Revenue : 매출 정보
- Competitors : 경쟁업체
- Easy Apply : 지원 난이도 (True, -1)



2. 모델 학습 전 데이터 최적화

Job Title	Data Analyst, Center on Immigration and Justice	Quality Data Analyst	Senior Data Analyst, Insights & Analytics Team
Salary Estimate	\$37K-\$66K (Glassdoor est.)	\$37K-\$66K (Glassdoor est.)	\$37K-\$66K (Glassdoor est.)
Job Description	Are you eager to roll up your sleeves and harness... Overview Provides analytical and technical ...		We're looking for a Senior Data Analyst who has... We're looking for a Senior Data Analyst who has... We're looking for a Senior Data Analyst who has...
Rating	3.2	3.8	3.4
Company Name	Vera Institute of Justice 3.2	Visiting Nurse Service of New York 3.8	Squarespace 3.4
Location	New York, NY	New York, NY	New York, NY
Headquarters	New York, NY	New York, NY	New York, NY
Size	201 to 500 employees	10000+ employees	1001 to 5000 employees
Founded	1961	1893	2003
Type of ownership	Nonprofit Organization	Nonprofit Organization	Company - Private
Industry	Social Assistance	Health Care Services & Hospitals	Internet
Sector	Non-Profit	Health Care	Information Technology
Revenue	\$100 to \$500 million (USD)	\$2 to \$5 billion (USD)	Unknown / Non-Applicable
Competitors	-1	-1	GoDaddy
Easy Apply	True	-1	-1



2. 모델 학습 전 데이터 최적화

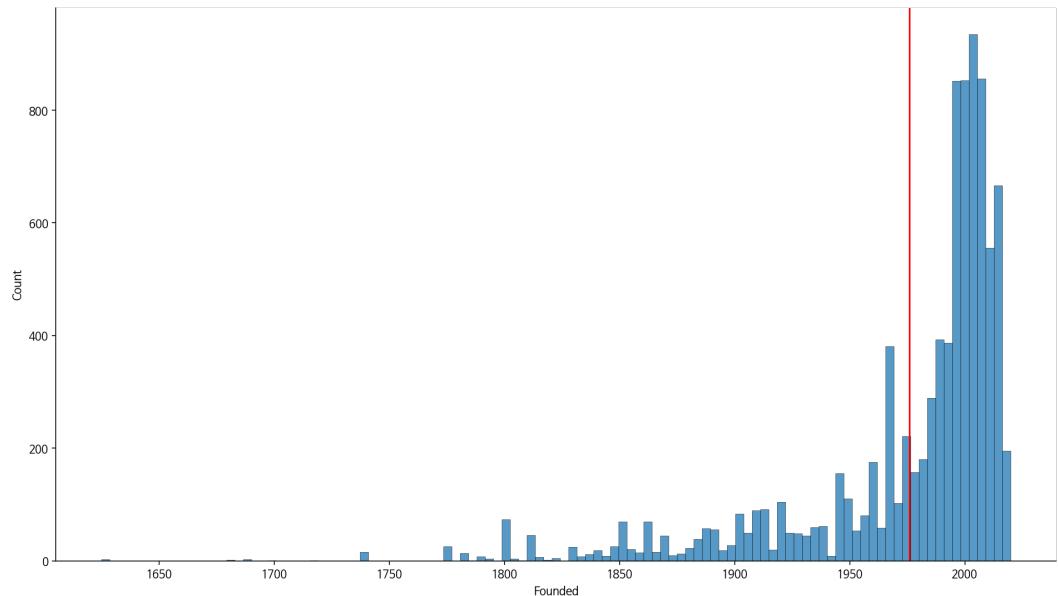
※ 데이터 전처리

순서	처리현황	비고
1	"-1"이 포함되어 있는 행 삭제 (Competitor, Easy Apply 제외)	전체 약 27%
2	Competitor : 경쟁사 수로 변경 Easy Apply : 0, 1 분류로 변경	-
3	직원 규모, 매출 → 등급으로 변경 연봉 → 평균으로 변경	직원 규모 : 1~7등급 매출 : 0~12등급
4	신규 특성 생성 본사=근무지일 경우 1, 아니면 0	특성 생성 Location=Headquarters
5	"Rating" 특성 삭제 → 신규 생성한 "Recommend"와 같은 정보의 데이터	데이터 누수 방지 머신 러닝 구축 시 적용

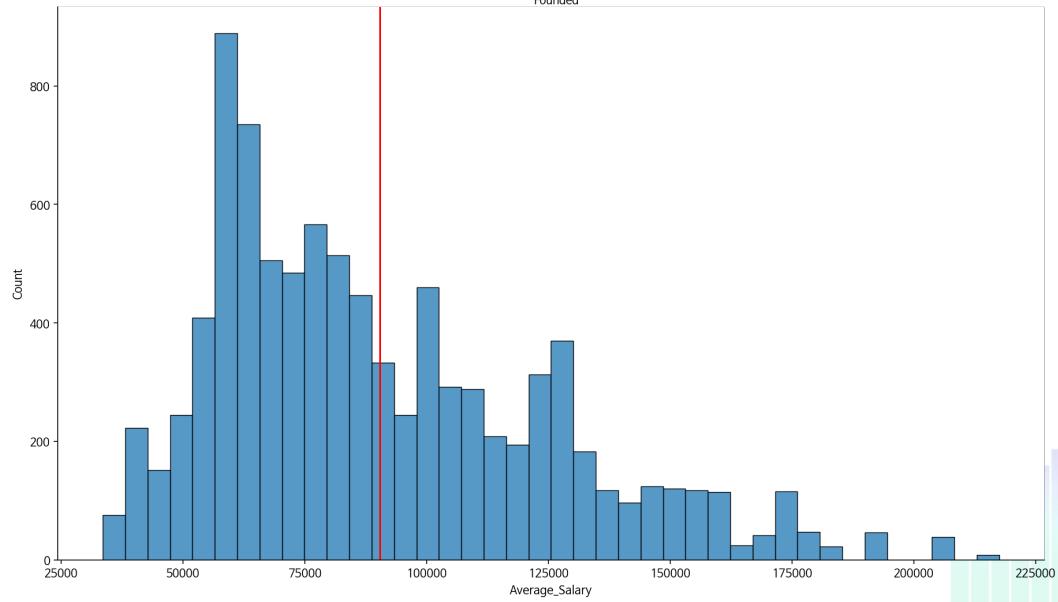


3. 데이터 시각화 및 요인 분석

설립 연도
평균
1976년



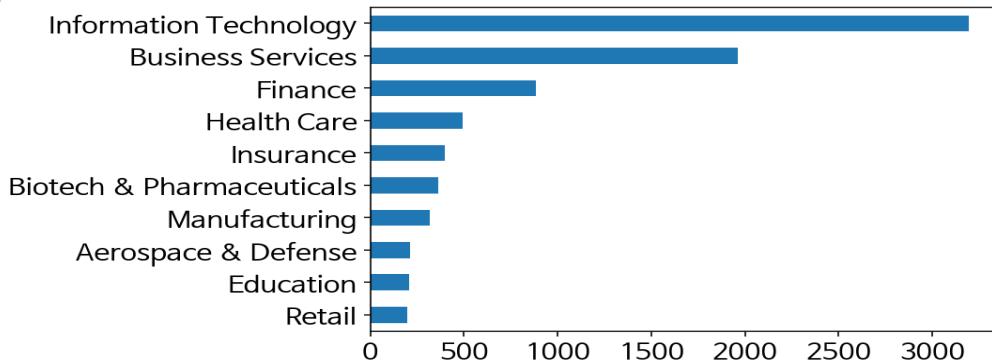
평균 연봉
약 9만불



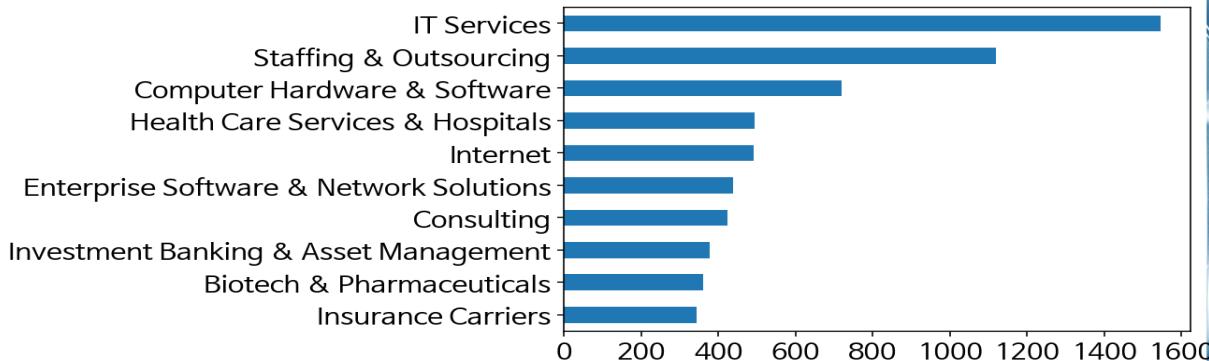
3. 데이터 시각화 및 요인 분석

※부문별 TOP10

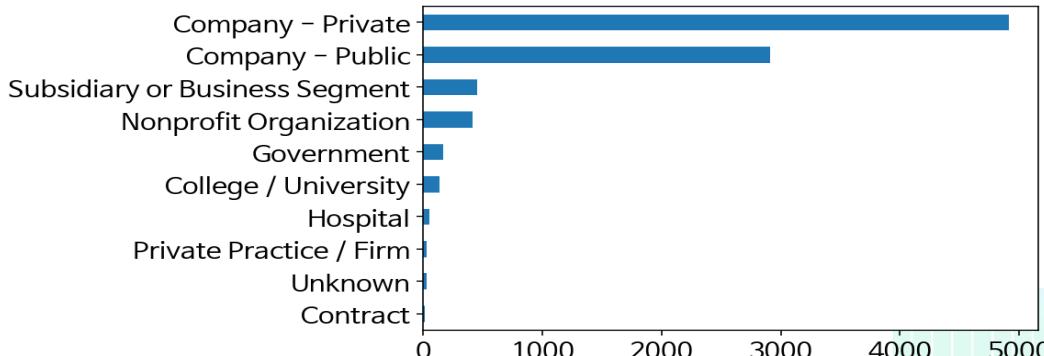
사업 부문



업종

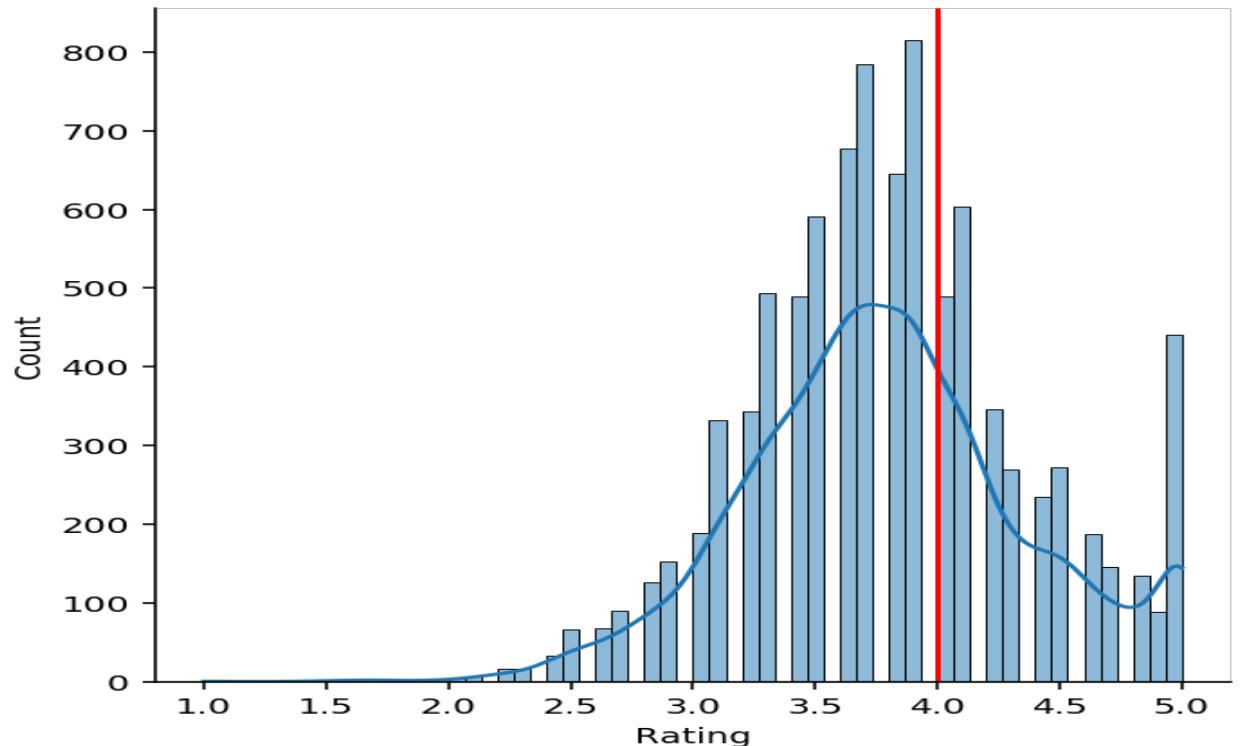


회사 유형



3. 데이터 시각화 및 요인 분석

※ 평점 분포

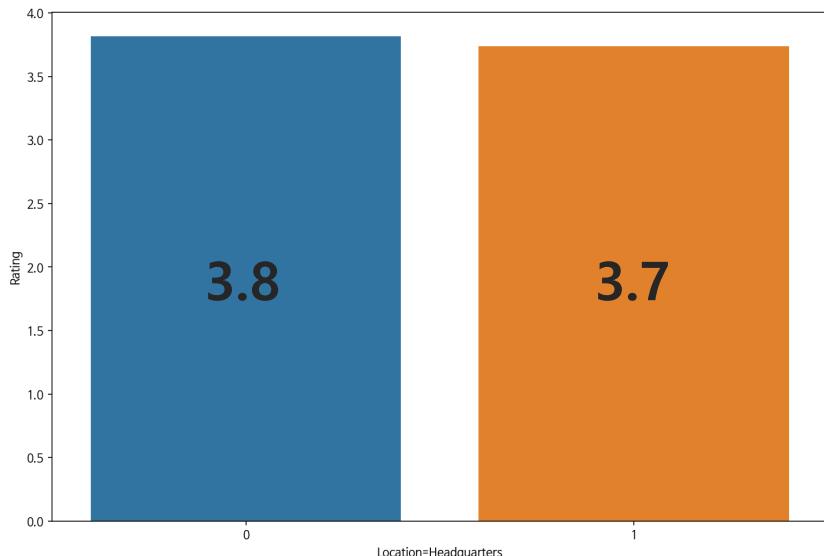
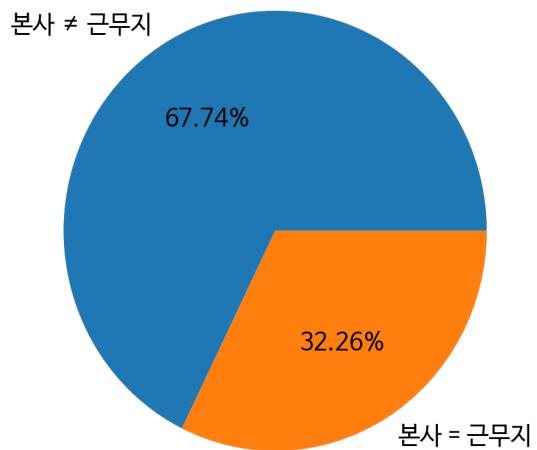


- 평균 평점 약 3.79



3. 데이터 시각화 및 요인 분석

※ 본사 / 지사 근무와 평점의 관계

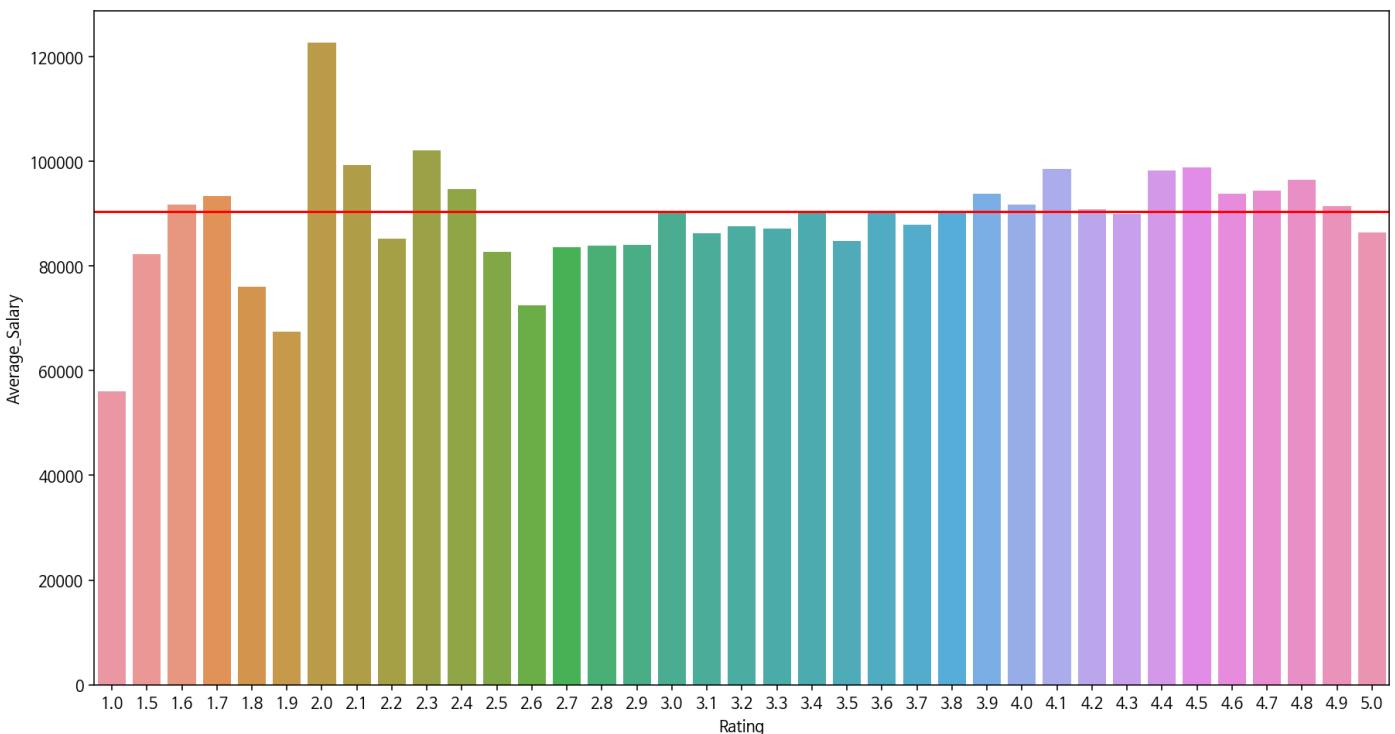


- 본사 근무 비율 약 32% / 지사 근무 비율 약 68%
- 본사 근무의 만족도(1), 지사 근무 만족도(0) 큰 차이는 없으나 지사 근무의 평균 만족도가 미세하게 높음



3. 데이터 시각화 및 요인 분석

※ 평점 별 평균 연봉



- 평점이 2.5 이하일 때 평점의 변동이 크지만 2.5 이상에서는 평균선에 근접하게 위치해 있음
→ 평점은 2.5~5.0 사이에 분포해 있기 때문에 2.5 이하의 표본이 적기 때문



3. 데이터 시각화 및 요인 분석

※ 회사 유형 / 업종 / 사업 부문 별 평균 평점

	Type of ownership	Rating	count%
0	College / University	4.1	1.5
1	Private Practice / Firm	3.9	0.36
2	Company - Private	3.9	53.64
3	Contract	3.8	0.2
4	Hospital	3.8	0.55
5	Company - Public	3.7	31.82
6	Subsidiary or Business Segment	3.6	4.99
7	Nonprofit Organization	3.6	4.55
8	Government	3.5	1.81
9	School / School District	3.4	0.19
10	Unknown	3.2	0.31
11	Other Organization	3.0	0.08
12	Self-employed	2.8	0.02

	Industry	Rating	count%
0	General Repair & Maintenance	4.7	0.01
1	Food Production	4.7	0.03
2	Music Production & Distribution	4.4	0.01
3	Colleges & Universities	4.2	1.7
4	IT Services	4.1	16.88
5	Grocery Stores & Supermarkets	4.0	0.26
6	Video Games	4.0	0.83
7	Internet	4.0	5.37
8	Enterprise Software & Network Solutions	3.9	4.78
9	Logistics & Supply Chain	3.9	0.48
10	General Merchandise & Superstores	3.9	0.07
11	Self-Storage Services	3.9	0.01
12	Computer Hardware & Software	3.9	7.86
13	Staffing & Outsourcing	3.9	12.22
14	Financial Analytics & Research	3.9	0.41
15	Research & Development	3.9	0.87
16	Transportation Equipment Manufacturing	3.9	0.53
17	Consulting	3.8	4.63
18	Motion Picture Production & Distribution	3.8	0.26
19	Education Training Services	3.8	0.23

	Sector	Rating	count%
0	Education	4.0	2.28
1	Information Technology	4.0	34.89
2	Business Services	3.8	21.42
3	Real Estate	3.8	0.48
4	Media	3.8	1.97
5	Accounting & Legal	3.8	0.94
6	Mining & Metals	3.7	0.01
7	Aerospace & Defense	3.7	2.31
8	Transportation & Logistics	3.6	0.92
9	Finance	3.6	9.63
10	Health Care	3.6	5.38
11	Oil, Gas, Energy & Utilities	3.6	1.42
12	Biotech & Pharmaceuticals	3.6	3.94
13	Arts, Entertainment & Recreation	3.6	0.25
14	Manufacturing	3.5	3.45
15	Government	3.5	1.8
16	Consumer Services	3.5	0.33
17	Travel & Tourism	3.5	0.09
18	Telecommunications	3.5	0.81
19	Agriculture & Forestry	3.5	0.1
20	Construction, Repair & Maintenance	3.5	0.33
21	Retail	3.5	2.15
22	Insurance	3.5	4.35
23	Restaurants, Bars & Food Services	3.2	0.14
24	Non-Profit	3.1	0.61

평균 평점 상위는 빈도가 적기 때문에 평균 평점이 높은 것으로 추측



3. 데이터 시각화 및 요인 분석

※ 시각화 및 데이터 분석에서 확인 결과

- 본사/지사 근무에 따른 평점 비교 큰 차이가 없음
- 평점 별 연봉 비교에서는 2.5 이하에서 큰 차이가 있으나 적은 표본으로 신뢰도 부족
- 회사 유형 / 업종 / 사업 부분 별 평점에서는 고평가 지표의 표본이 적기 때문에 신뢰도 부족으로 판단

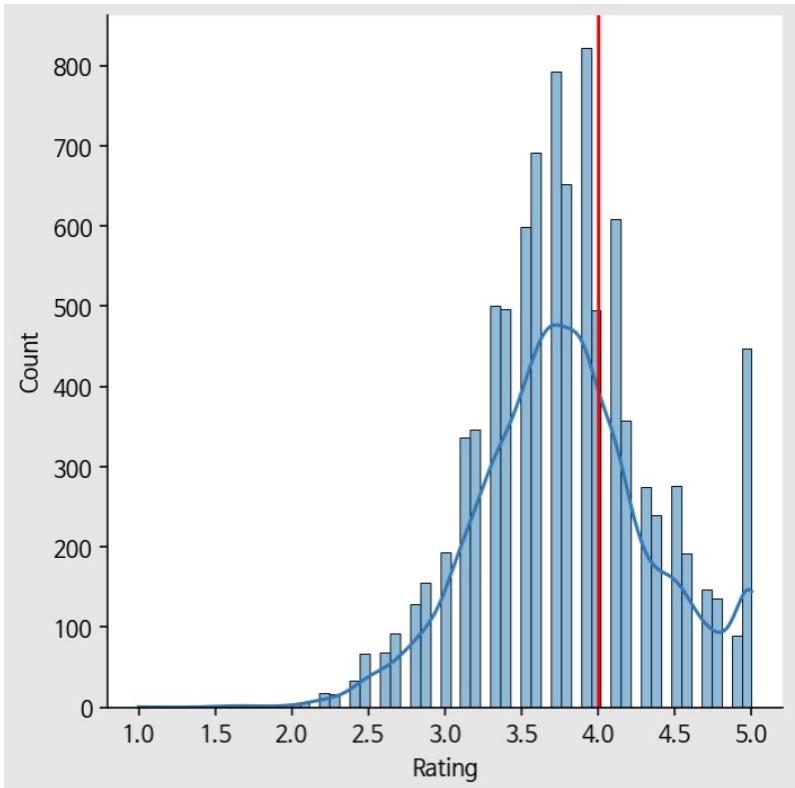
※ 머신러닝을 활용한 분석

- 시각화 분석이 아닌 머신러닝 활용
- 특정 평점 기준 만족/불만족을 나는 이진분류 모델을 통해 만족/불만족에 큰 영향이 있는 특성을 찾아볼 예정
- 모델비교 의사결정나무, 랜덤 포레스트 분류기, XGBoost 분류기



4. 머신러닝을 활용한 분석

※ 기준 모델 : Rating 4.0이상 추천 (이진 분류 모델)



- “Recommend” 4.0 이상 1, 4.0 이하 0 인 신규 특성 생성
- 전체의 35%가 추천하는 것으로 나타남
- 기준 모델의 정확도 : 0.65



4. 머신러닝을 활용한 분석

※ 모델 평가

- 학습 / 검증 / 테스트(20%) 세트 구성
- 모델 학습 및 검증 절차를 거쳐 20%데이터로 최종 결과 산출

※ 3가지 분류모델 학습 결과 (검증 데이터)

Rank	모델	정확도	F1 점수	정밀도	재현율
1	XGBoost 분류기	0.86	0.79	0.80	0.78
2	의사결정나무	0.83	0.77	0.73	0.82
3	랜덤 포레스트 분류기	0.81	0.73	0.68	0.81

※ XGBoost 분류기모델 교차 검증

- 학습 데이터를 3등분하여 모델 성능 평가
- 정확도 : 0.801, 0.790, 0.799
- f1 점수 : 0.679, 0.734, 0.692

→ 정확도 & f1의 점수차는 나지만, 점수는 균등하게 나옴



4. 머신러닝을 활용한 분석

※ 최종 모델 평가

구분	정확도	F1 점수	정밀도	재현율	AUC
훈련	0.91	0.86	-	-	-
검증	0.86	0.79	0.80	0.78	0.92
테스트	0.84	0.76	0.80	0.73	0.91

- 정확도 : 훈련/검증 결과보다 하락하였으나, 기준 모델 0.65 보다 향상 되었음
- 해당 모델의 경우 “정밀도>재현율”이 좋은 모델
 - 1) 직업에 대해 실제 만족하면서 비추천
 - 2) 직업에 대해 실제 불만족하면서 추천
- AUC : 0.91

※ 정확도 : 정확하게 예측한 비율

※ 정밀도 : 모델이 True 라고 분류한 것 중 실제 True인 비율

※ 재현율 : 실제 True 인것중 모델이 True라고 예측한 비율

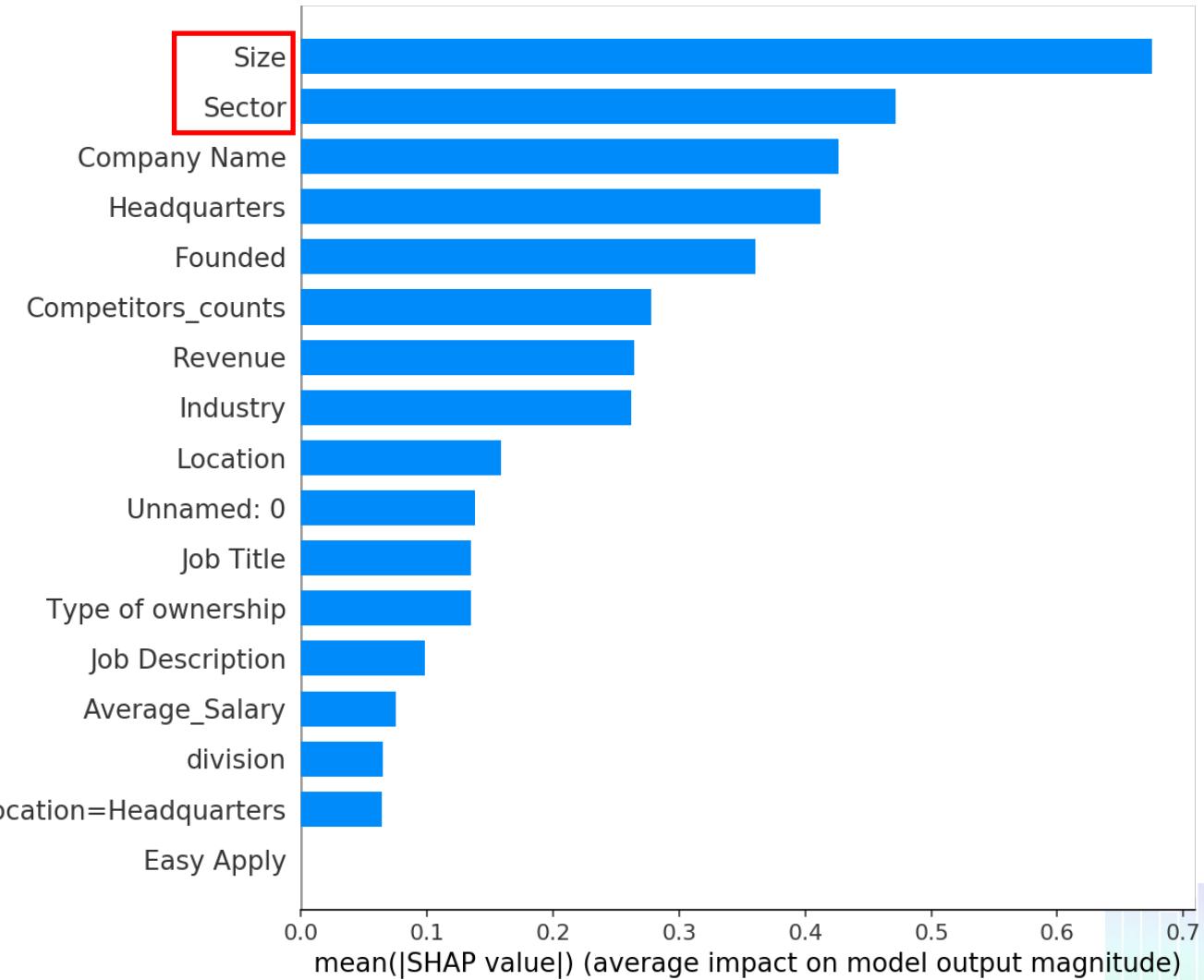
※ f1 점수 : 정밀도와 재현율을 조화평균

※ AUC점수 : 모델이 예측을 잘하는지 측정하는 평가지표, 1에 가까울 수록 좋은 모델



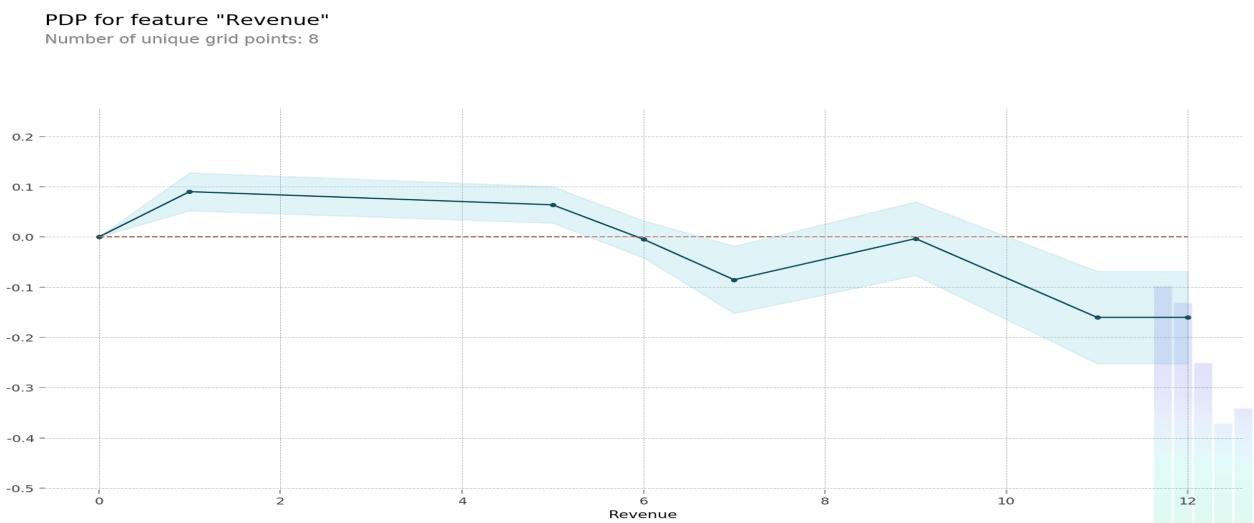
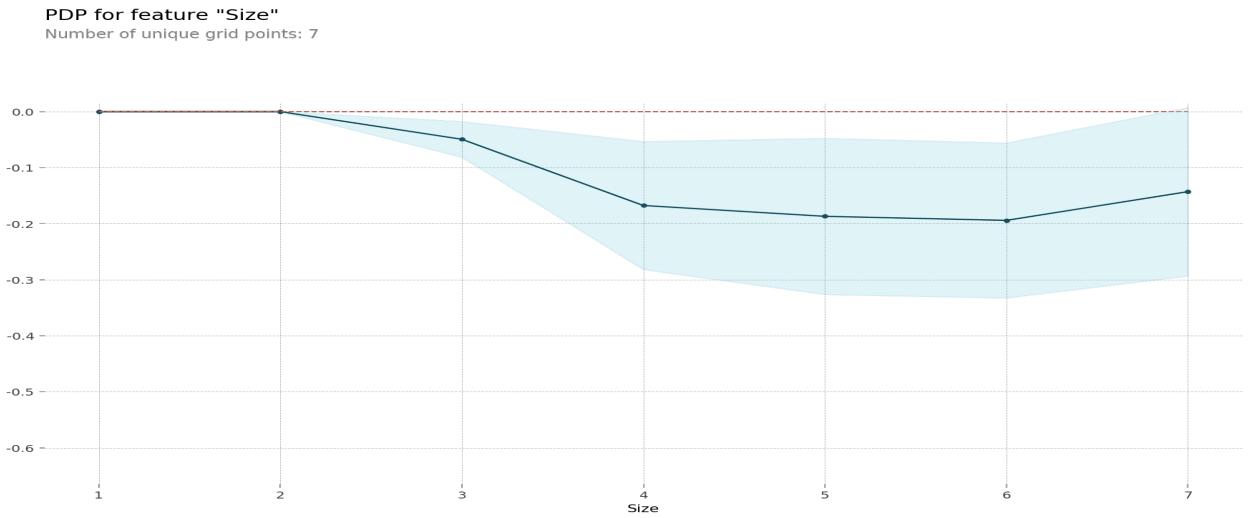
4. 머신러닝을 활용한 분석

※ 특성들의 중요도



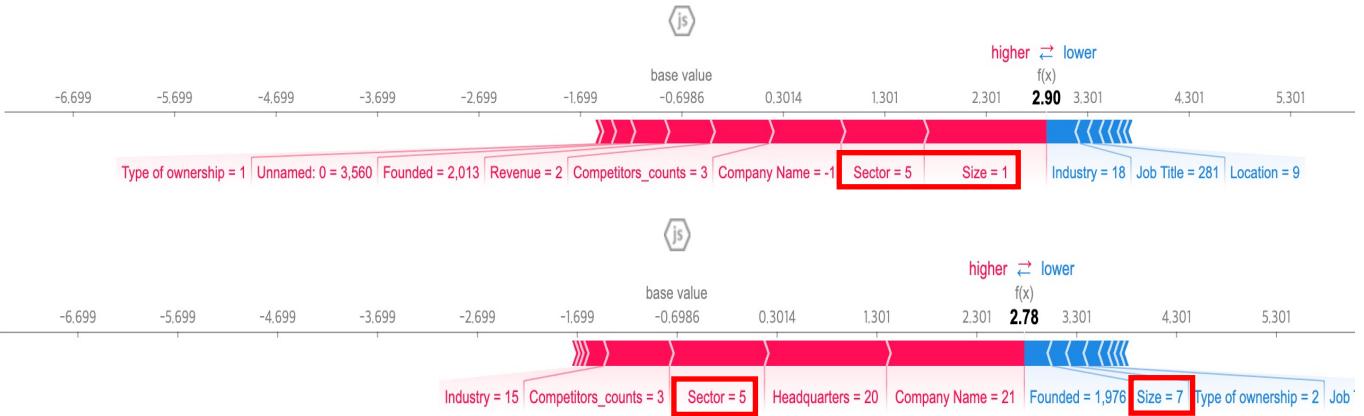
4. 머신러닝을 활용한 분석

※ 결과에 대한 직원 수 & 매출의 영향

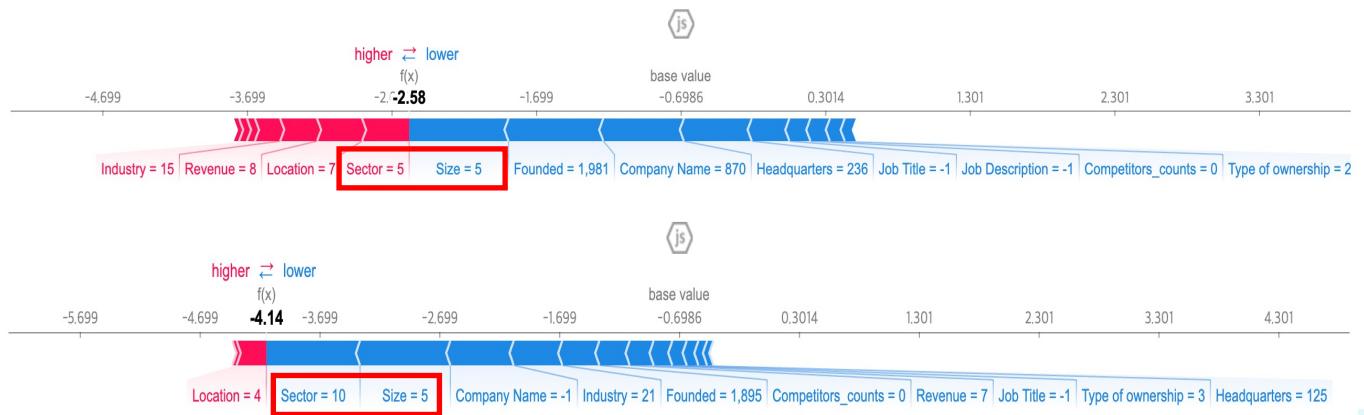


4. 머신러닝을 활용한 분석

※ Recommend : 1(만족) 샘플 분석



※ Recommend : 0(불만족) 샘플 분석



4. 머신러닝을 활용한 분석

※ 결과

- 이진 분류 모델의 특성 중요도에서는 인원규모(Size)와 사업부문(Sector)가 상위 2개 특성으로 확인 되었음
- 인원규모가 증가 할 수록 만족도에 대해 부정적인 요인이 된다는 것을 확인 할 수 있음
- 실제 데이터 만족2건, 불만족 2건에 대해 shap을 통해 확인해 본 결과 인원규모가 작을 수록 만족도에 주요 영향으로 볼 수 있고 클 수록 부정적인 요인으로 확인됨
- 사업부문의 경우 인코딩이 되어 정확히 특정할 순 없으나 만족도에 영향력이 높은 것으로 판단 됨



5. 프로젝트 인사이트 및 회고

※ 알게 된 것

- 시각화 분석으로는 어느 특성이 평점에 대한 큰 요인인지 확인하기 어려웠으나, 머신러닝 모델을 통하여 상세 요인을 확인할 수 있었음
- 블랙박스 형태인 머신러닝에 shap라이브러리를 활용하여 상세 분석 가능하다는 것을 확인 하였음

※ 추후 보완 사항

- 사람인, 잡코리아 등 국내 자료로 한국에서는 어떤 요인이 데이터 직군의 만족도가 높은지에 대해 분석 해볼 수 있음
- 이번 프로젝트에서는 카테고리 특성에 대한 상관관계 분석을 진행하지 않았지만, 추후에는 카이제곱검정, 아노바검정 등을 활용하여 분석 진행
- 해당 결과로 시각화 비교를 위해 대쉬보드를 만들어 볼 수 있음



Thank you

