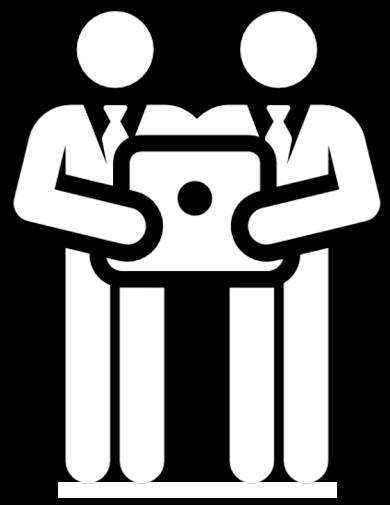
Are they candidates for work at the company?

AIB Section 2 Project

노주연 (jooyeonroh@gmail.com)

2022.05.24



CONTENTS

1. 프로젝트 개요

- 데이터셋 소개
- 프로젝트 목표
- 데이터 가설 및 평가지표

2. 데이터 전처리

- baseline Model
- Feature Engineering
- 가설 검증

3. 머신러닝 모델링

- Decision Tree
- Random Forest
- XGBoost

4. 최종모델

데이터셋

데이터 소개

지원자 아이디, 도시코드, 도시개발지수, 성별, 관련 경험 여부, 대학교 과정, 교육수준, 전공분야, 경력, 현재 소속된 회사규모, 회사 유형, 이직기간, 훈련 시간

데이터 크기

(19158, 14)

프로젝트 목표

목표

수집한 지원자 데이터를 통해 **지원자가 직업 전환을 희망하는지 여부**를 예측하는 머신러닝 모델을 완성한다.

타겟

0 – 직업 전환 희망하지 않음
1 – 직업 전환 희망

가설 설정

가설 1

오랜 경력을 가진 지원자는 직업 전환을 희망하지 않을 것이다.

- 경력직들이 직업전환을 원하지 않는다면, 경력직들만을 위한 훈련 과정을 개설하여, 회사의 발전에 큰 기여할 수 있는 직원을 채용할 수 있다.

가설 2

적은 시간 훈련을 받은 지원자들은 다른 직업으로 전환한다.

- 적은 시간만 훈련을 받은 지원자들이 다른 직업으로 전환한다는 가설이 검증된다면, 채용에 최소 훈련 시간 기준을 세워 회사에서 올바른 채용 결정을 할 수 있다.

결측값

	NA_cout	NA_ratio
enrollee_id	0	0.00
city	0	0.00
city_development_index	0	0.00
gender	4508	0.24
relevent_experience	0	0.00
enrolled_university	386	0.02
education_level	460	0.02
major_discipline	2813	0.15
experience	65	0.00
company_size	5938	0.31
company_type	6140	0.32
last_new_job	423	0.02
training_hours	0	0.00

방법1.
데이터의 비율대로 결측치 채우기

방법2.
결측값 제거

방법3.
'other' 혹은 'unknown' 값으로 채우기

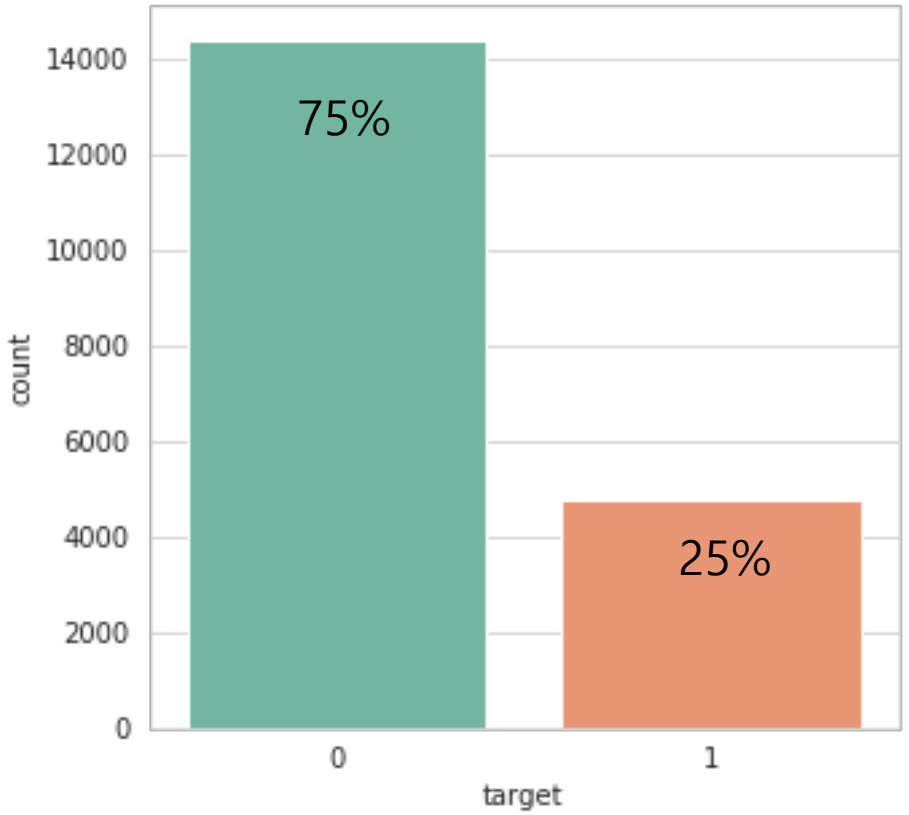
Baseline Model

베이스라인 모델이란?

- 모델의 성능을 비교하기 위한 초기 모델

- 최빈값을 통해 직업전환을 희망하지 않는 지원자는 75%임을 확인

타겟 분포도



가설 검증1

오랜 경력을 가진 지원자는 직업 전환을 희망하지 않을 것이다.

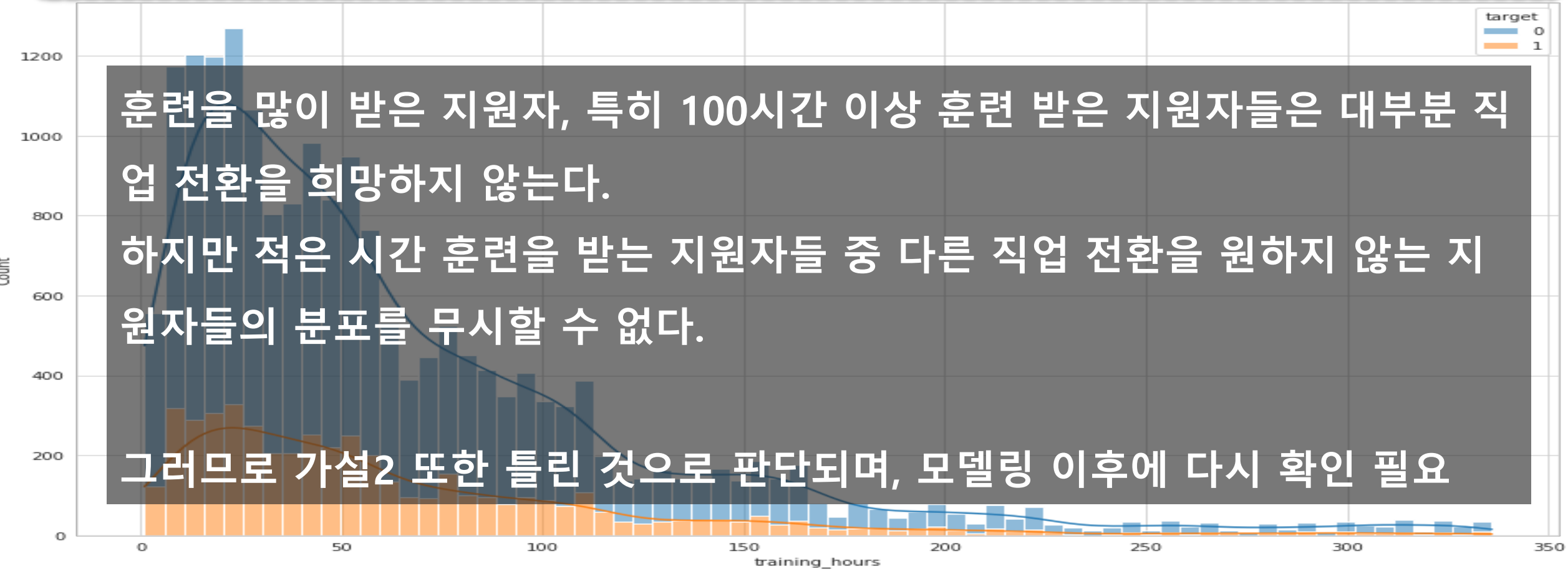
경력이 20년 이상 된 지원자는 눈에 띄게 직업 전환을 희망하지 않았다.

하지만, 경력 연수에 따라 직업 전환을 희망하는지 아닌지 예측하는 것은 어렵다.

그러므로 가설은 틀린 것으로 판단 되며, 모델링 이후에 다시 확인 필요

가설 검증2

적은 시간 훈련을 받은 지원자들은 다른 직업으로 전환한다.

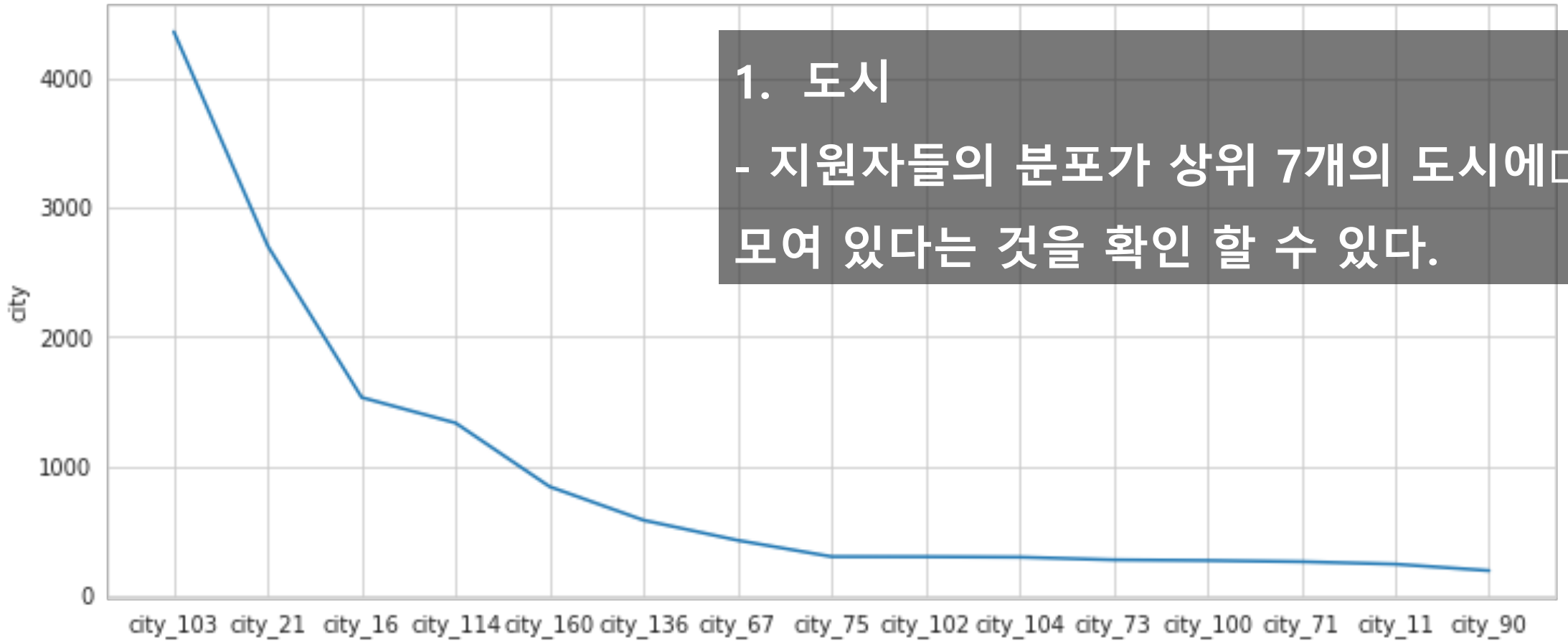


훈련을 많이 받은 지원자, 특히 100시간 이상 훈련 받은 지원자들은 대부분 직업 전환을 희망하지 않는다.

하지만 적은 시간 훈련을 받는 지원자들 중 다른 직업 전환을 원하지 않는 지원자들의 분포를 무시할 수 없다.

그러므로 가설2 또한 틀린 것으로 판단되며, 모델링 이후에 다시 확인 필요

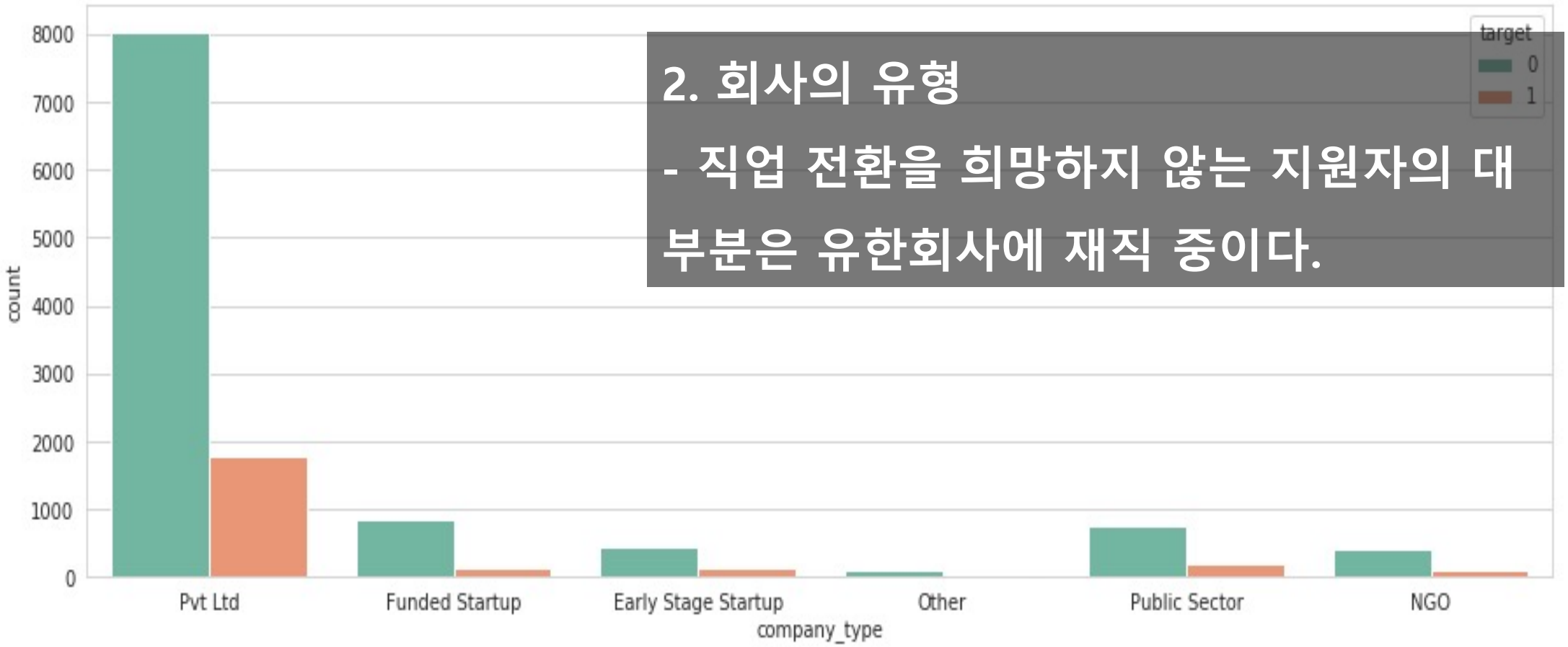
EDA



1. 도시

- 지원자들의 분포가 상위 7개의 도시에 모여 있다는 것을 확인 할 수 있다.

EDA



2. 회사의 유형

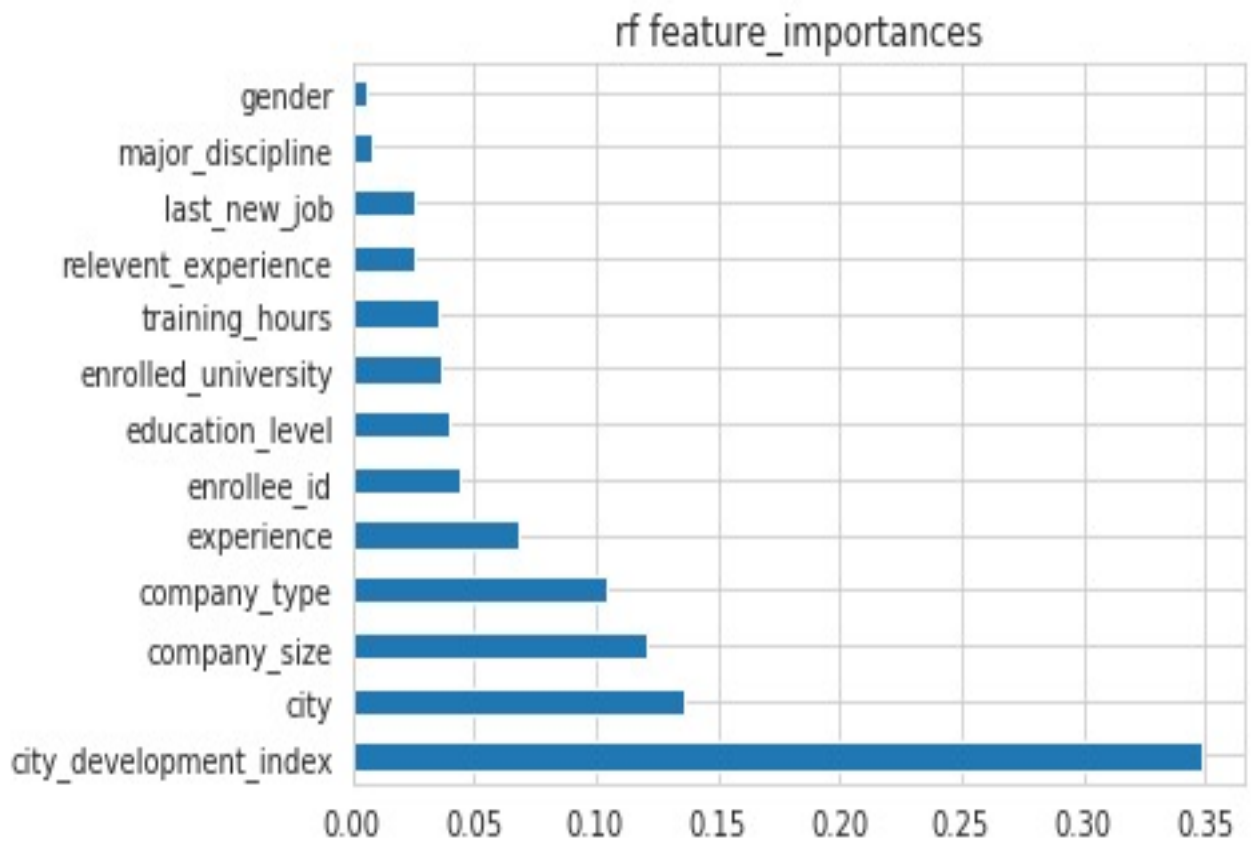
- 직업 전환을 희망하지 않는 지원자의 대부분은 유한회사에 재직 중이다.

Random Forest

성능평가

Random Forest Score	
Accuracy	0.785409
Recall	0.313368
Precision	0.630017
ROC AUC Score	**0.626587**

feature_importances

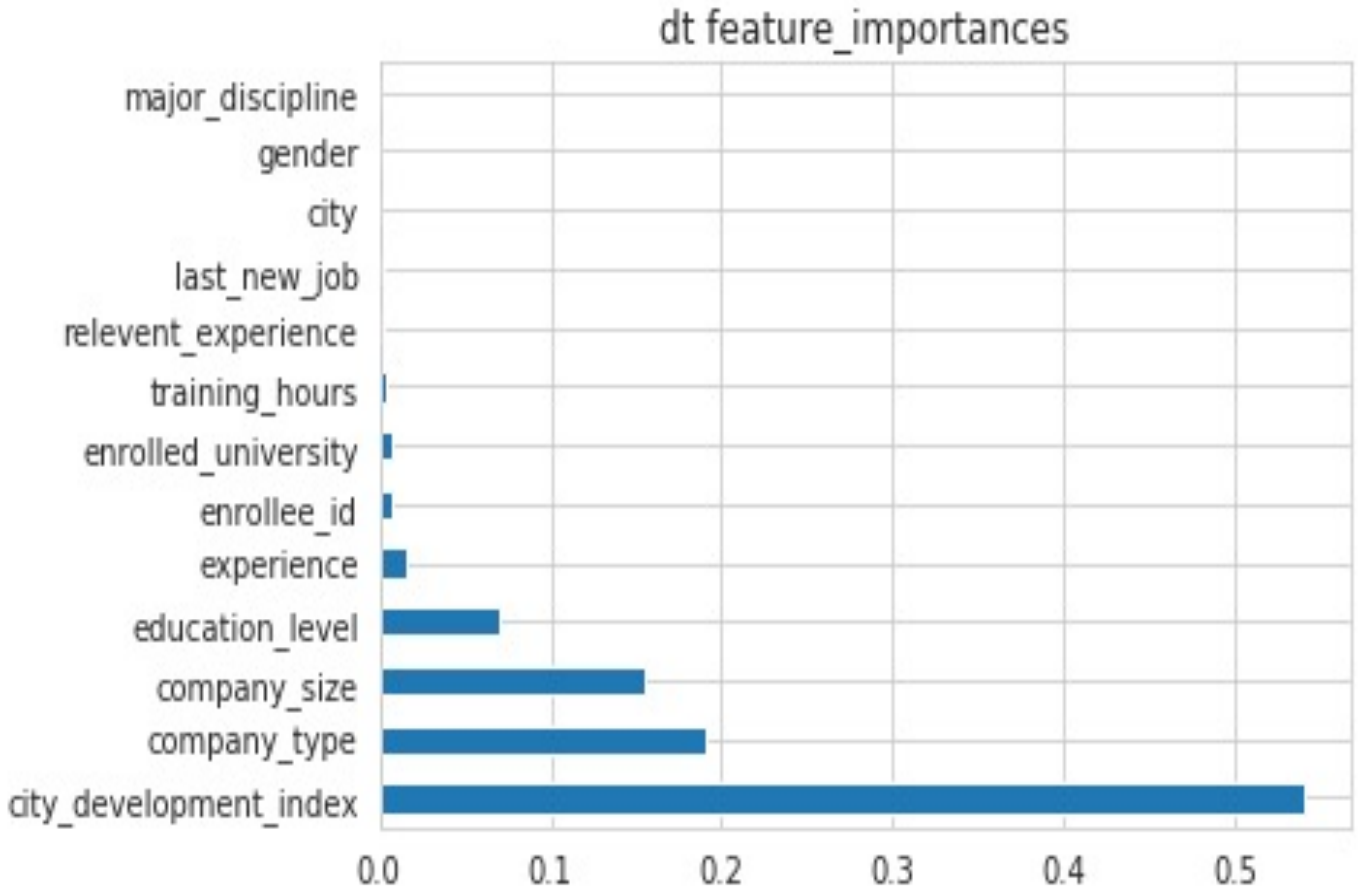


Decision Tree

성능평가

Decision Tree Score	
Accuracy	0.790329
Recall	0.387153
Precision	0.619444
ROC AUC Score	0.654678

feature_importances

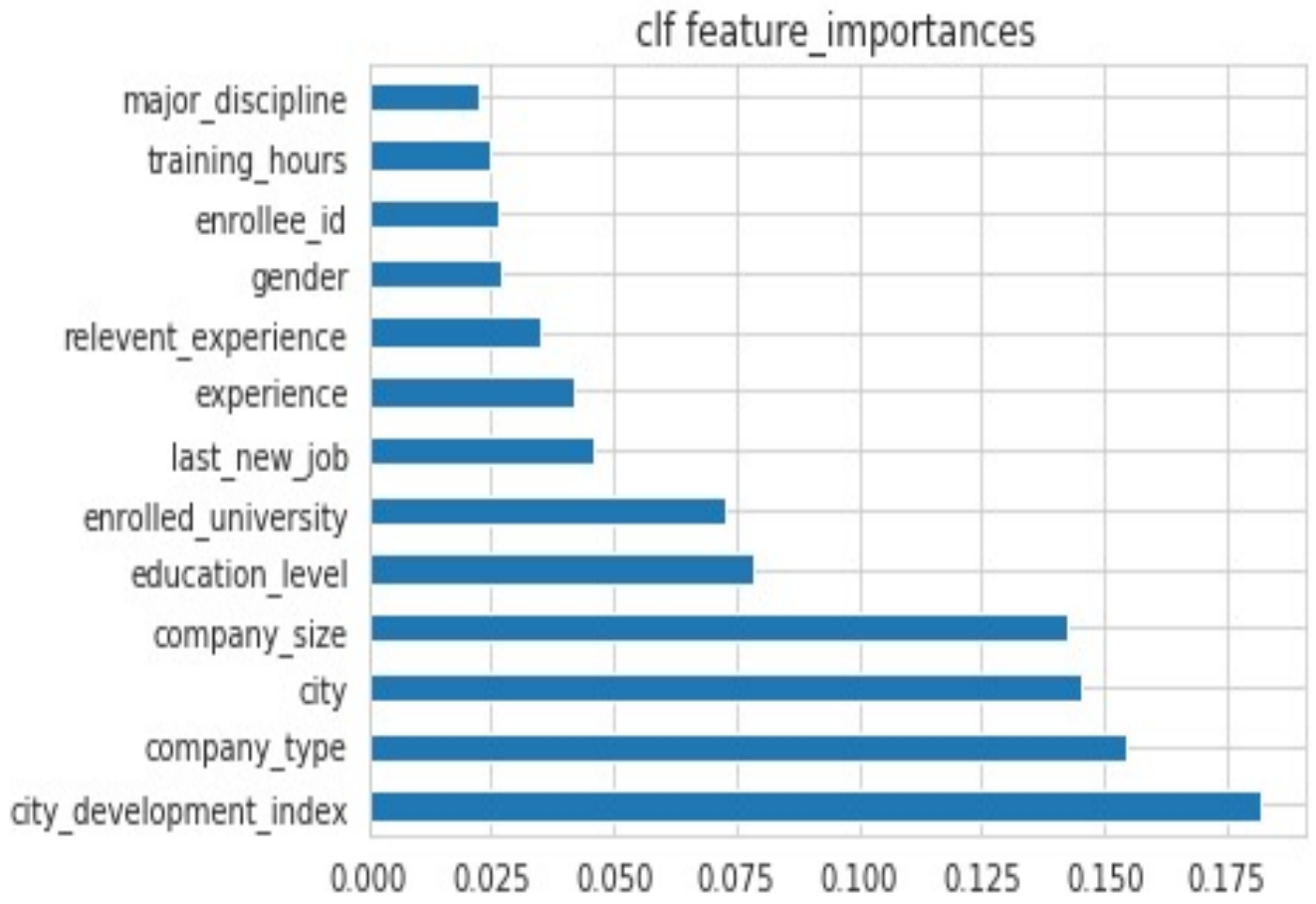


XGBoost

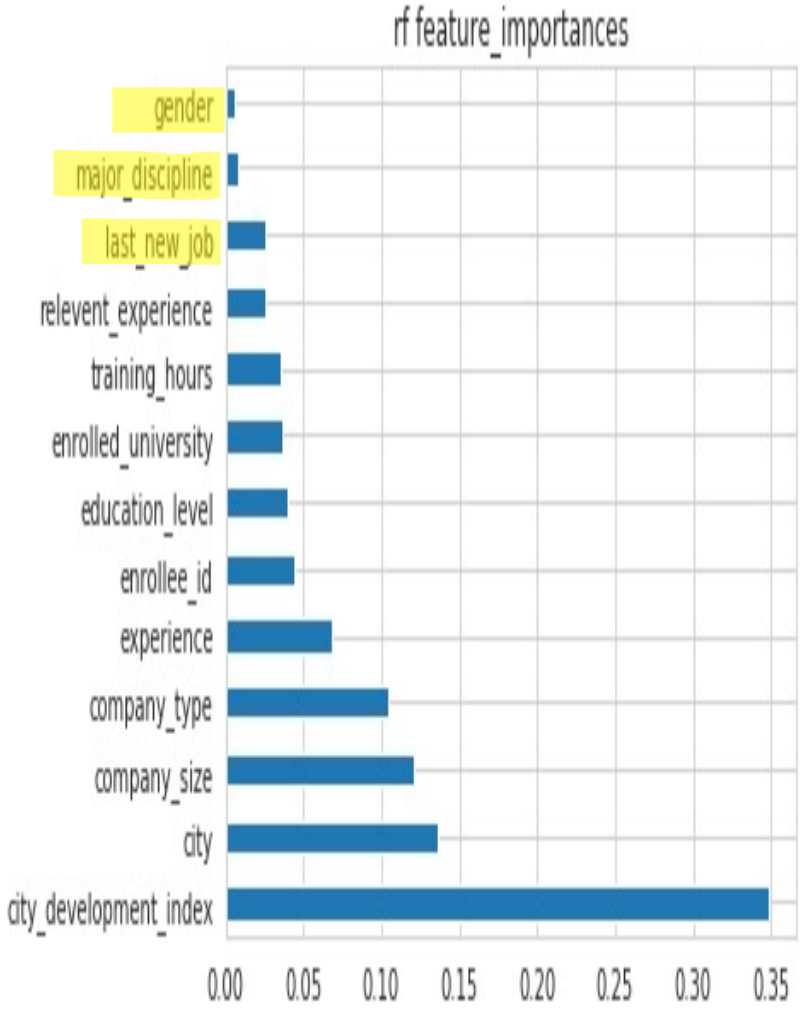
성능평가

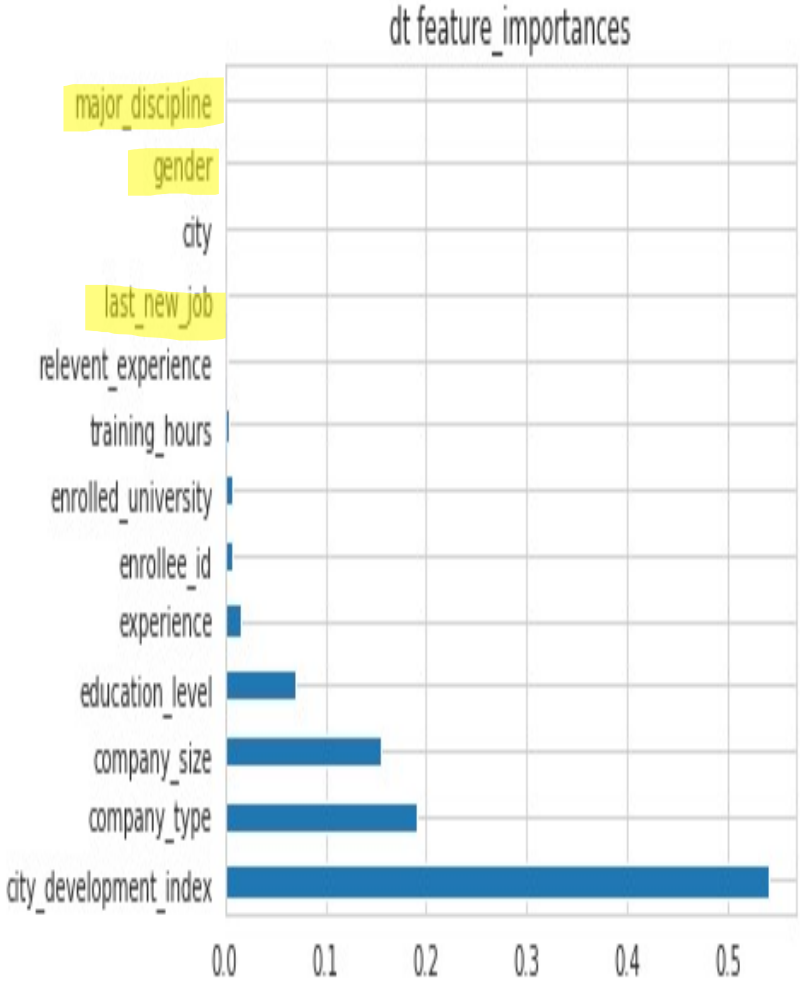
Decision Tree Score	
Accuracy	0.785623
Recall	0.353299
Precision	0.612952
ROC AUC Score	0.640164

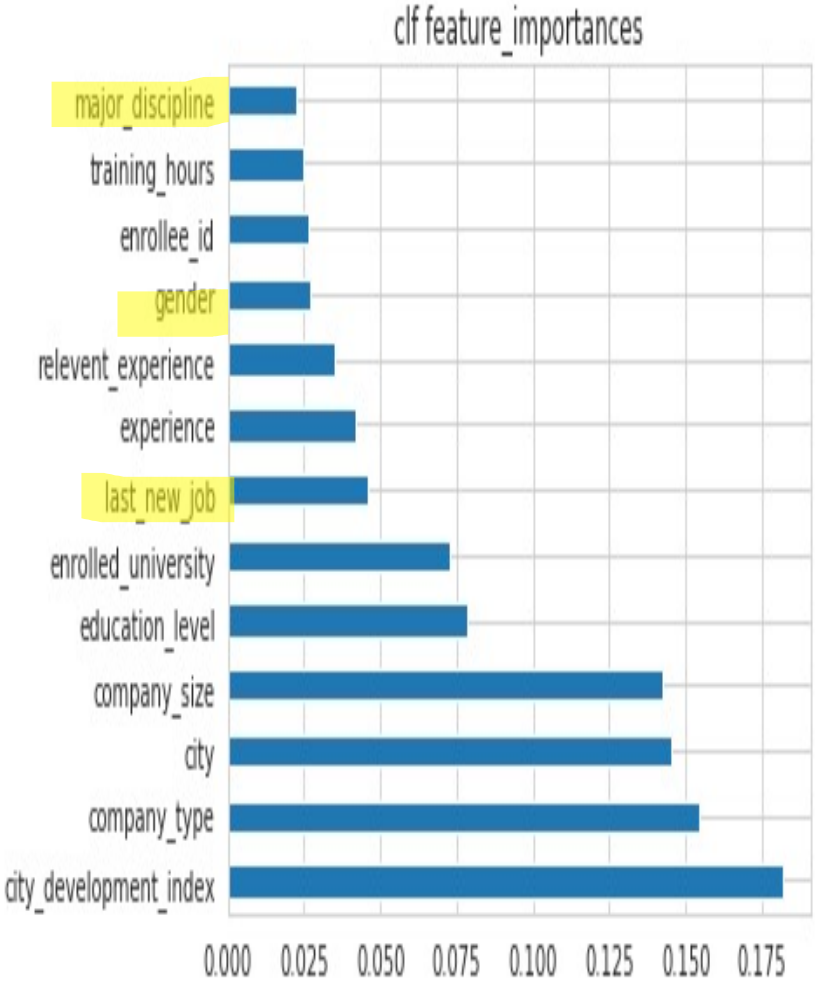
feature_importances



특성 중요도 비교


rf feature_importances
gender
major_discipline
last_new_job
relevent_experience
training_hours
enrolled_university
education_level
enrollee_id
experience
company_type
company_size
city
city_development_index
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35


dt feature_importances
major_discipline
gender
city
last_new_job
relevent_experience
training_hours
enrolled_university
enrollee_id
experience
education_level
company_size
company_type
city_development_index
0.0
0.1
0.2
0.3
0.4
0.5


clf feature_importances
major_discipline
training_hours
enrollee_id
gender
relevent_experience
experience
last_new_job
enrolled_university
education_level
company_size
city
company_type
city_development_index
0.000
0.025
0.050
0.075
0.100
0.125
0.150
0.175

최종 모델 : Random Forest

최종 모델 성능평가

Random Forest Score	
Accuracy	0.789688
Recall	0.375000
Precision	0.621583
ROC AUC Score	**0.650163**

Weight	Feature
0.0396 ± 0.0083	city_development_index
0.0284 ± 0.0031	city
0.0102 ± 0.0044	experience
0.0009 ± 0.0010	training_hours
0.0005 ± 0.0012	relevent_experience
0.0003 ± 0.0028	enrollee_id
0 ± 0.0000	company_type
0 ± 0.0000	company_size
0 ± 0.0000	education_level
0 ± 0.0000	enrolled_university

곧 ! 돌아옵니다!

모델링을 실패하여 의미 있는
인사이트를 얻을 수 없었습니다.

리모델링 중으로 추후 업데이트 될 예정입니다.

더 멋진 모습으로 찾아 뵙겠습니다 !