직원 이탈 여부 예측

in India

AI_15_신희호

발표 순서

1. 문제 정의
2. 가설 설정
3. 데이터 설명
4. 데이터 전처리
5. 탐색적 데이터 분석
6. 모델링 및 모델 결과
7. 결론

1. 문제 정의

문제 정의

어떤 특성을 가진 직원이 이탈할지 예측하고,
이를 바탕으로 직원의 이탈을 막기 위해
어떤 전략을 세우면 좋을지 살펴보고자 합니다.

1. 직원의 이탈에 영향을 가장 많이 주는 특성은 무엇일까요?
2. 직원의 이탈을 막기 위해 세울 수 있는 전략은 무엇일까요?

2. 가설 설정

가설 설정

1. 최근 입사한 직원일수록 이탈 비율이 높을 것이다.

2. 경험이 적은 직원일수록 이탈 비율이 높을 것이다.

3. 나이가 어린 직원일수록 이탈 비율이 높을 것이다.

3. 데이터 설명

데이터 설명

1. Education - 학위 : Bachelors(학사), Masters(석사), PHD(박사)
2. JoiningYear - 입사 연도 : 2012~2018년
3. City - 근무지 : Bangalore(Bengaluru), Pune, New Delhi
4. PaymentTier - 급여 수준 : 1, 2, 3 (1이 가장 높고, 3이 가장 낮음)
5. Age - 나이 : 22~41세
6. Gender - 성별 : Male(남자), Female(여자)
7. EverBenched - 프로젝트에 참가하지 못했는지 여부 : Yes(참가하지 못함), No(참가했음)
8. ExperienceInCurrentDomain - 현재 도메인에서 경력 : 0~7년
9. LeaveOrNot(타겟) - 조직을 떠날 것인지 여부 : 0(떠나지 않음), 1(떠남)

4. 데이터 전처리

데이터 전처리

1. 결측치 없음

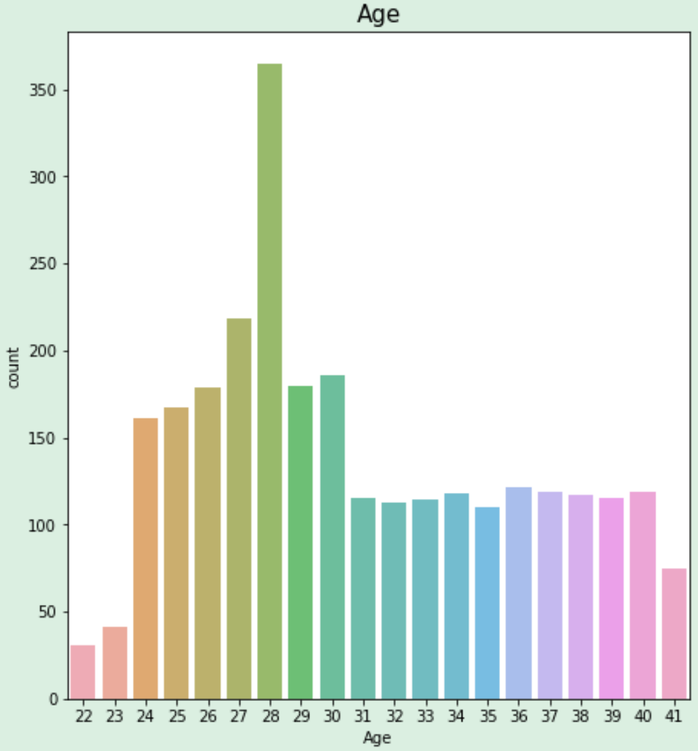
2. 이상치 없음

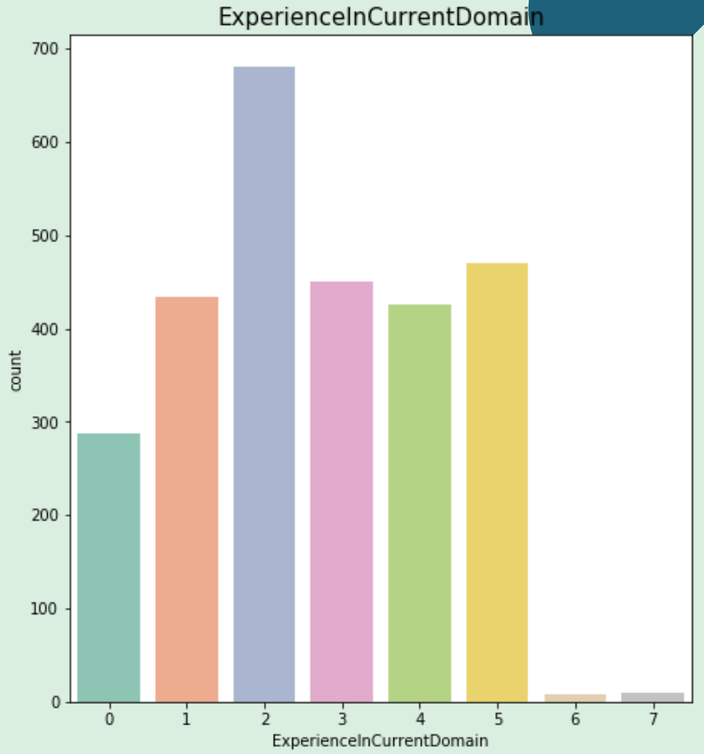
3. 중복행 1889건 발견
   > 드랍 결정

4. Age 그룹화
   > 22-25 = 1, 26-29 = 2,
   30-33 = 3, 34-37 = 4,
   38-41 = 5

5. ExperienceInCurrentDomain
   > 6, 7년 데이터 수 부족으로 드랍
   > 특성 이름을 Experience로 간소화

6. Education
   > Bachelors = 1, Masters = 2, PHD = 3으로 변경





5. 탐색적 데이터 분석

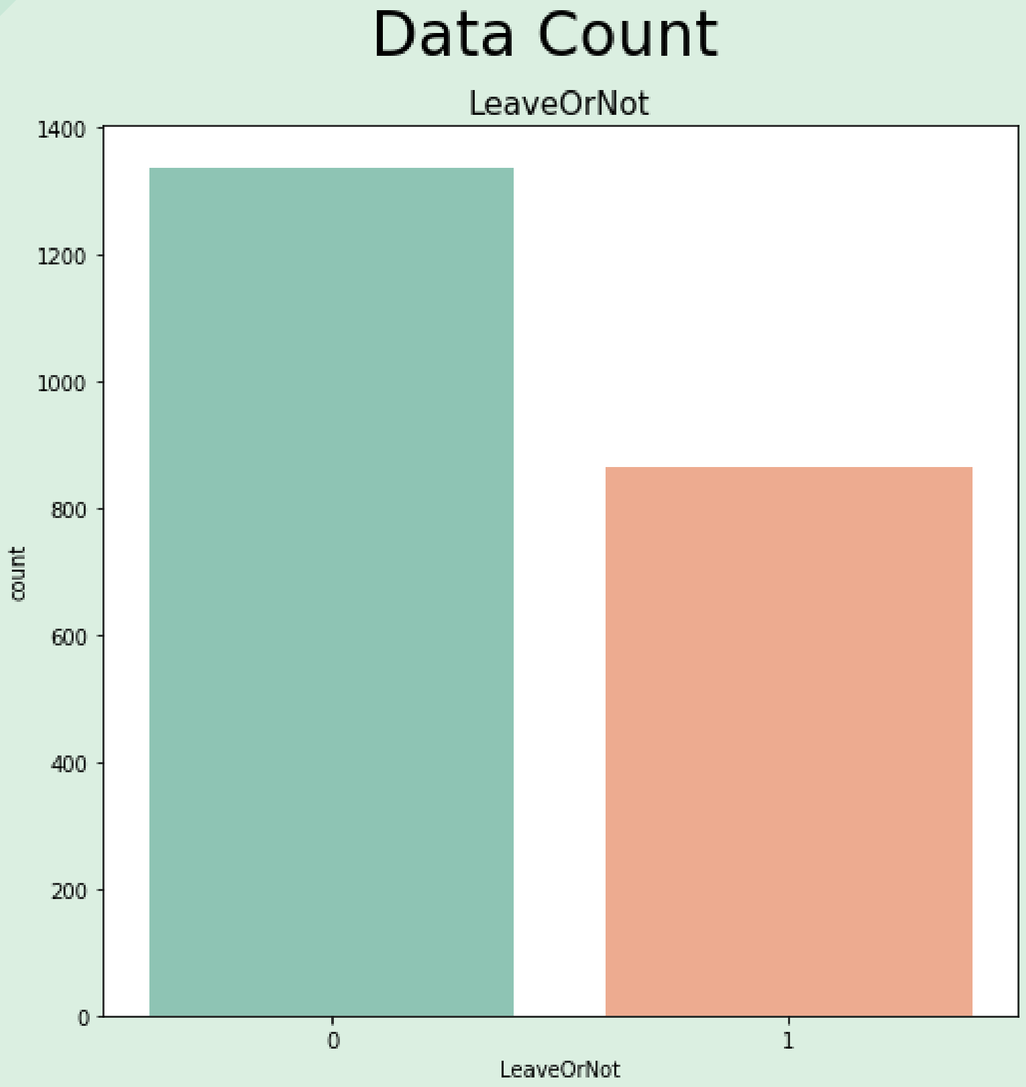
탐색적 데이터 분석

1. Train, Test 세트로 분리
   > Train 0.8, Test 0.2 비율로 분리
   > Train 2199행, Test 548행

2. Train, Test 타겟 비율 확인
   > Train의 타겟 비율 : 0.39
   > Test의 타겟 비율 : 0.40
   > 0과 1의 불균형이 심하지 않고,
   > Train, Test 사이의 불균형도 심하지 않는 것으로 확인됩니다.

Test는 예측해야 하는 데이터이므로 모른다고 가정해야 합니다.
따라서, Train만 탐색적 데이터 분석을 진행합니다.

탐색적 데이터 분석


Data Count
LeaveOrNot
1400
1200
1000
800
600
400
200
0
count
0
1
LeaveOrNot

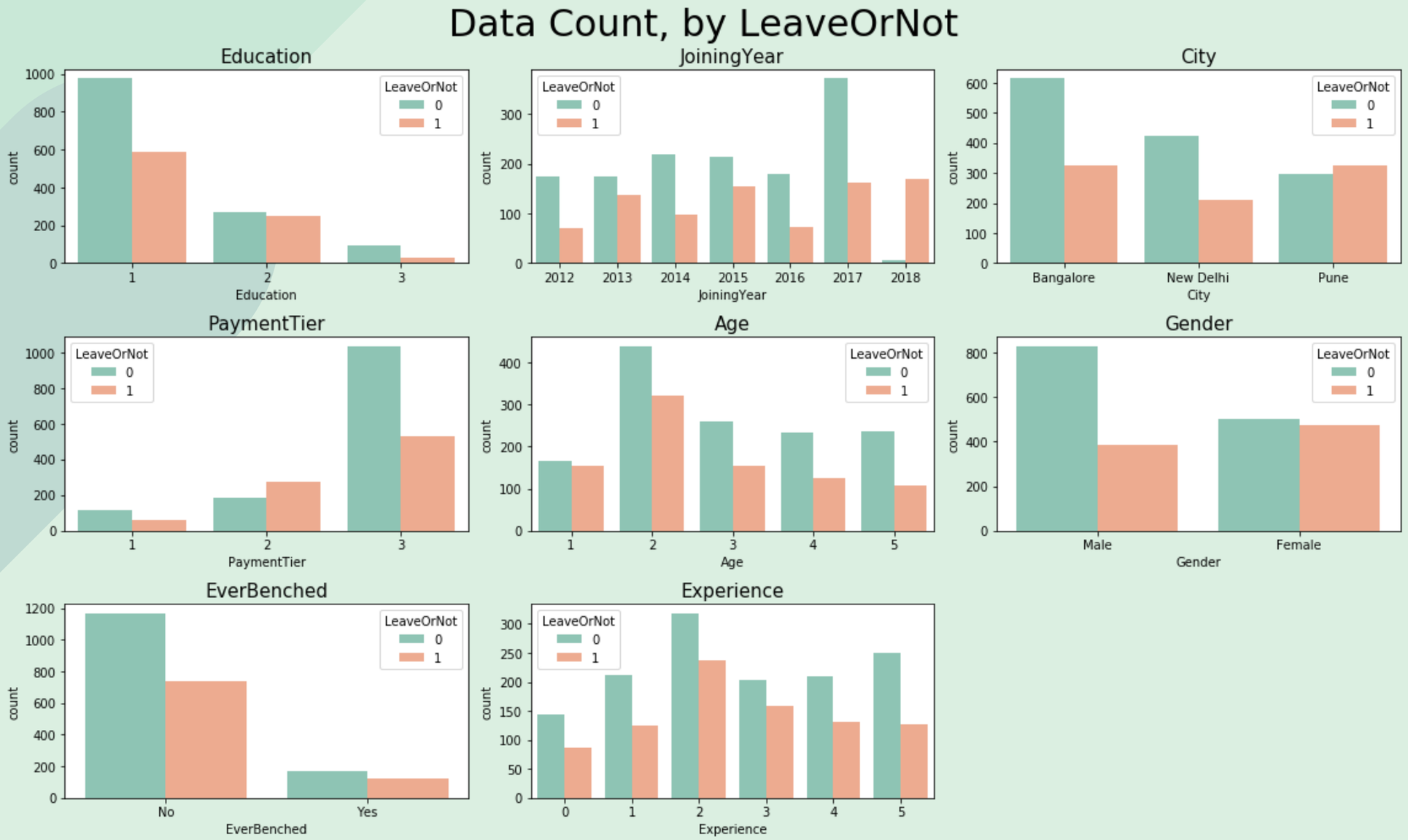
탐색적 데이터 분석


Data Count
Education
count
Education
JoiningYear
count
JoiningYear
City
count
City
Bangalore
New Delhi
Pune
PaymentTier
count
PaymentTier
Age
count
Age
Gender
count
Male
Female
Gender
EverBenched
count
No
Yes
EverBenched
Experience
count
Experience

탐색적 데이터 분석


Feature Distributions
Education
JoiningYear
City
PaymentTier
Age
Gender
EverBenched
Experience
LeaveOrNot
Bangalore
New Delhi
Pune
Male
Female
No
Yes

탐색적 데이터 분석


Data Count, by LeaveOrNot
Education
JoiningYear
City
PaymentTier
Age
Gender
EverBenched
Experience
LeaveOrNot
0
1
count
Education
JoiningYear
City
PaymentTier
Age
Gender
EverBenched
Experience
Bangalore
New Delhi
Pune
Male
Female
No
Yes

탐색적 데이터 분석


Data Rate
Education
LeaveOrNot
Education
JoiningYear
JoiningYear
City
Bangalore
New Delhi
Pune
City
PaymentTier
PaymentTier
Age
Age
Gender
Male
Female
Gender
EverBenched
No
Yes
EverBenched
Experience
Experience

탐색적 데이터 분석


LeaveOrNot
0.013
0.13
-0.12
-0.11
-0.021
1
Education
JoiningYear
PaymentTier
Age
Experience
LeaveOrNot
1.0
0.8
0.6
0.4
0.2
0.0

6. 모델링 및 모델 평가

모델링

1. 기준 모델을 통한 기본 성능 파악
   > 이진 분류 모델이므로 최빈값으로 기준 모델 설정
   > 정확도 값으로 0.61

2. 트리 모델을 사용하기로 하고, 우선 비교할 모델링
   > AdaBoost, XGBoost, LightGBM 선정

3. 범주형 변수(City, Gender, EverBenched)
   > Ordinal Encoding

4. 정확도와 F1을 성능 지표로 사용
   > 정확도는 기준 모델과의 비교를 위해 사용
   > F1은 모델 간 비교를 위해 사용

5. KFold를 통해 성능 측정

모델링

모델 간 비교

```
XGBoost_정확도 : 0.7945
XGBoost_F1 : 0.6893
```

```
LightGBM_정확도 : 0.7853
LightGBM_F1 : 0.6922
```

```
AdaBoost_정확도 : 0.7549
AdaBoost_F1 : 0.6256
```

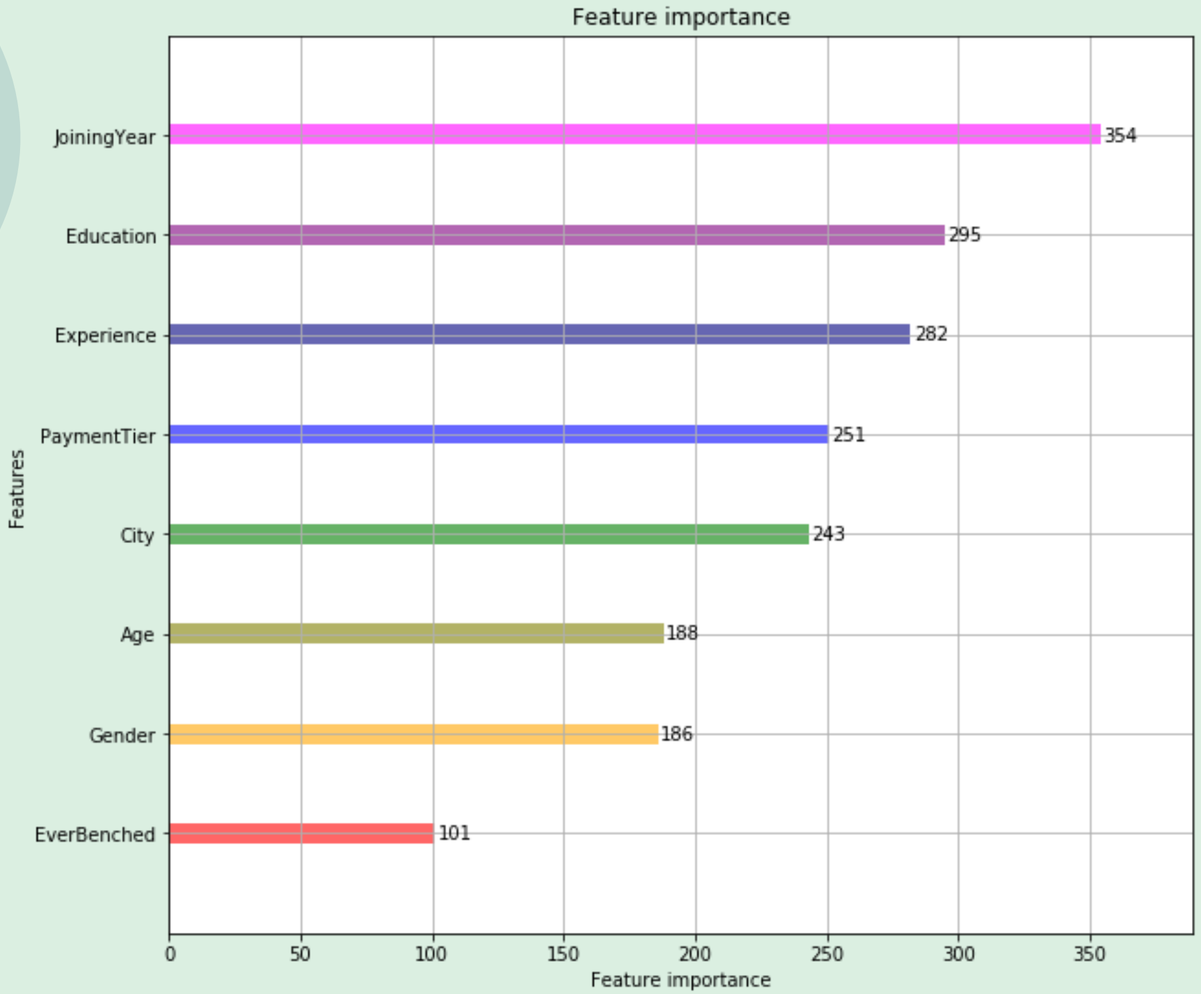
LightGBM 조정 후

```
LightGBM_정확도 : 0.8327
LightGBM_F1 : 0.7468
```

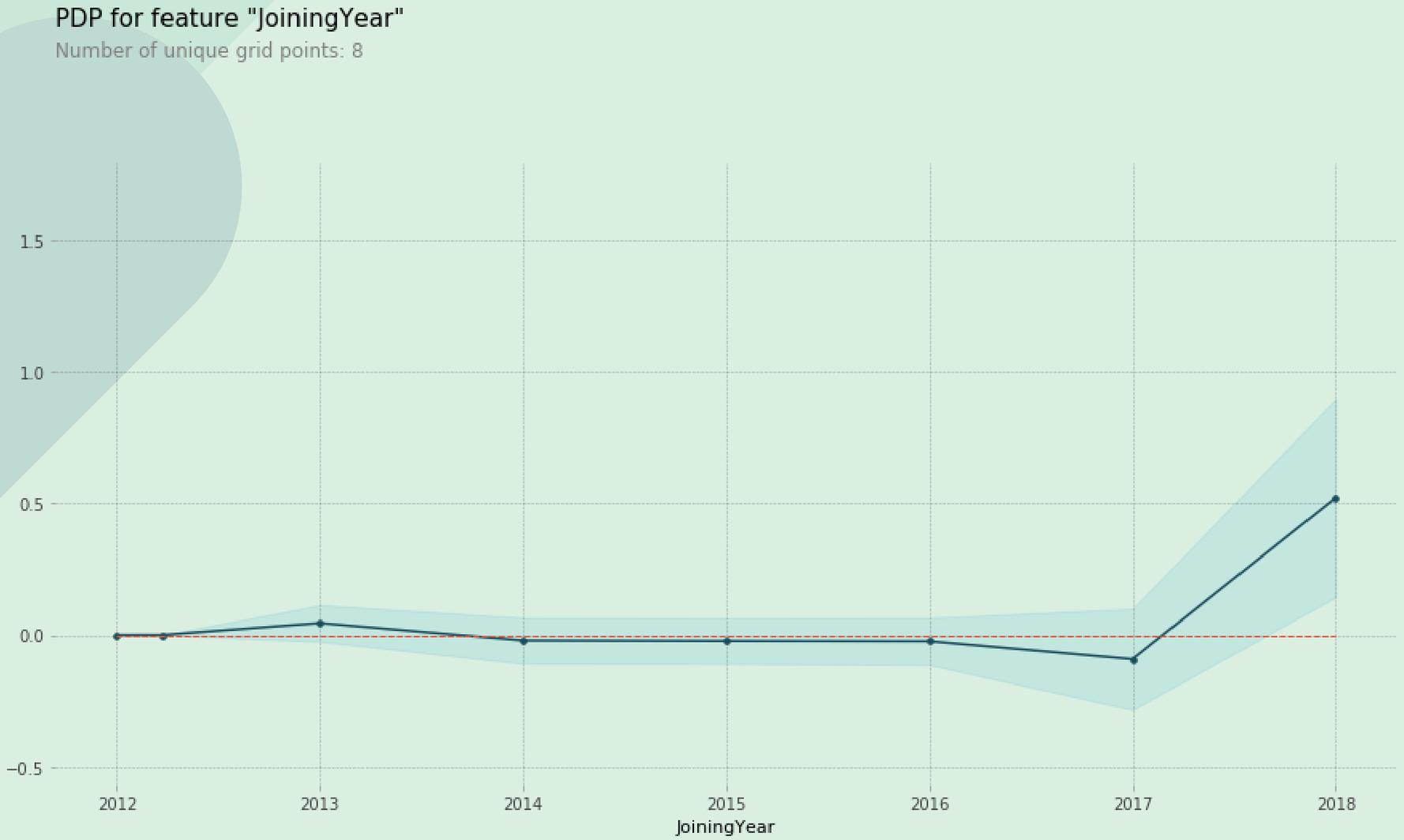
LightGBM 조정 후 Test 세트 결과

```
LightGBM_정확도 : 0.823
LightGBM_F1 : 0.7413
```

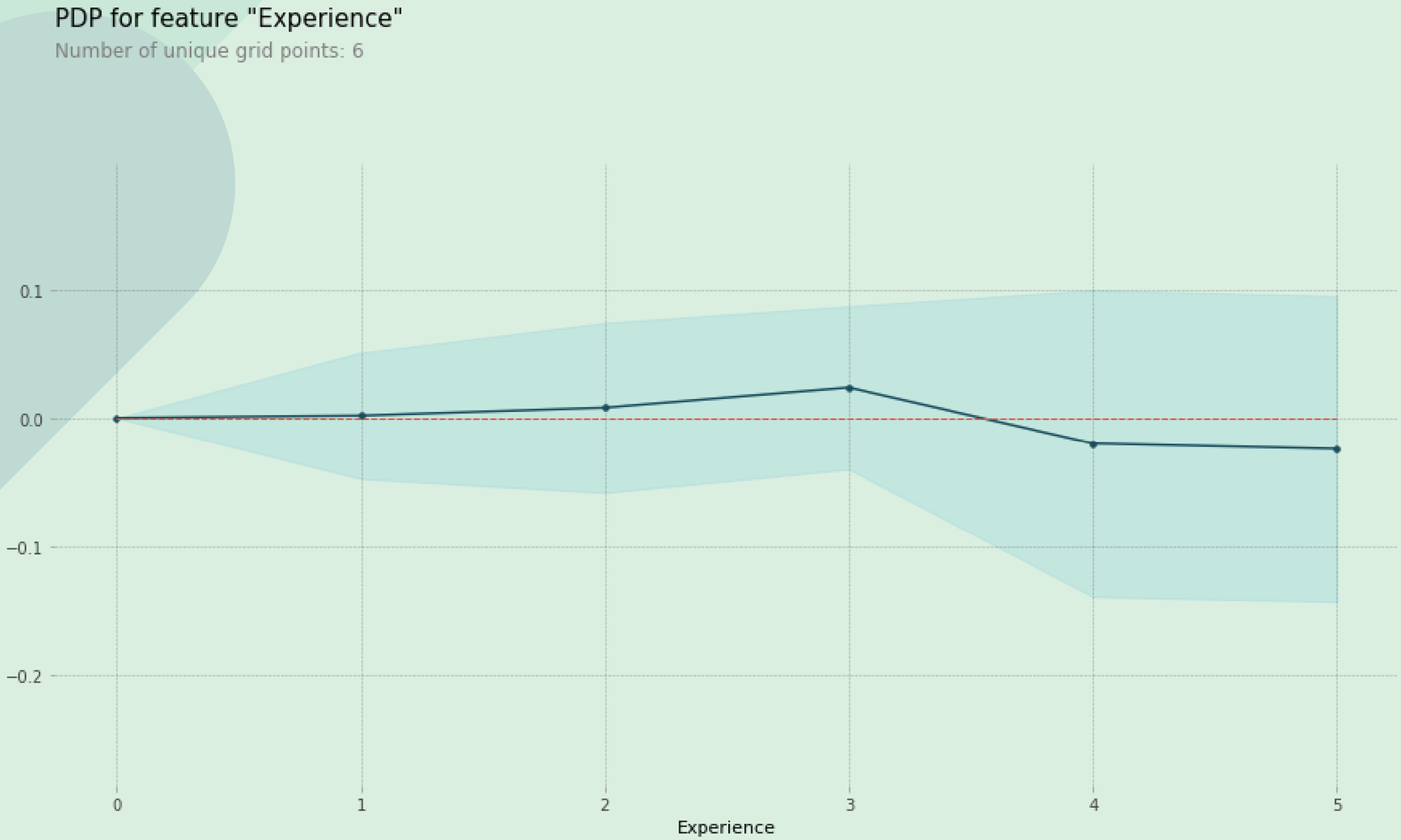
특성 중요도(Feature Importance)



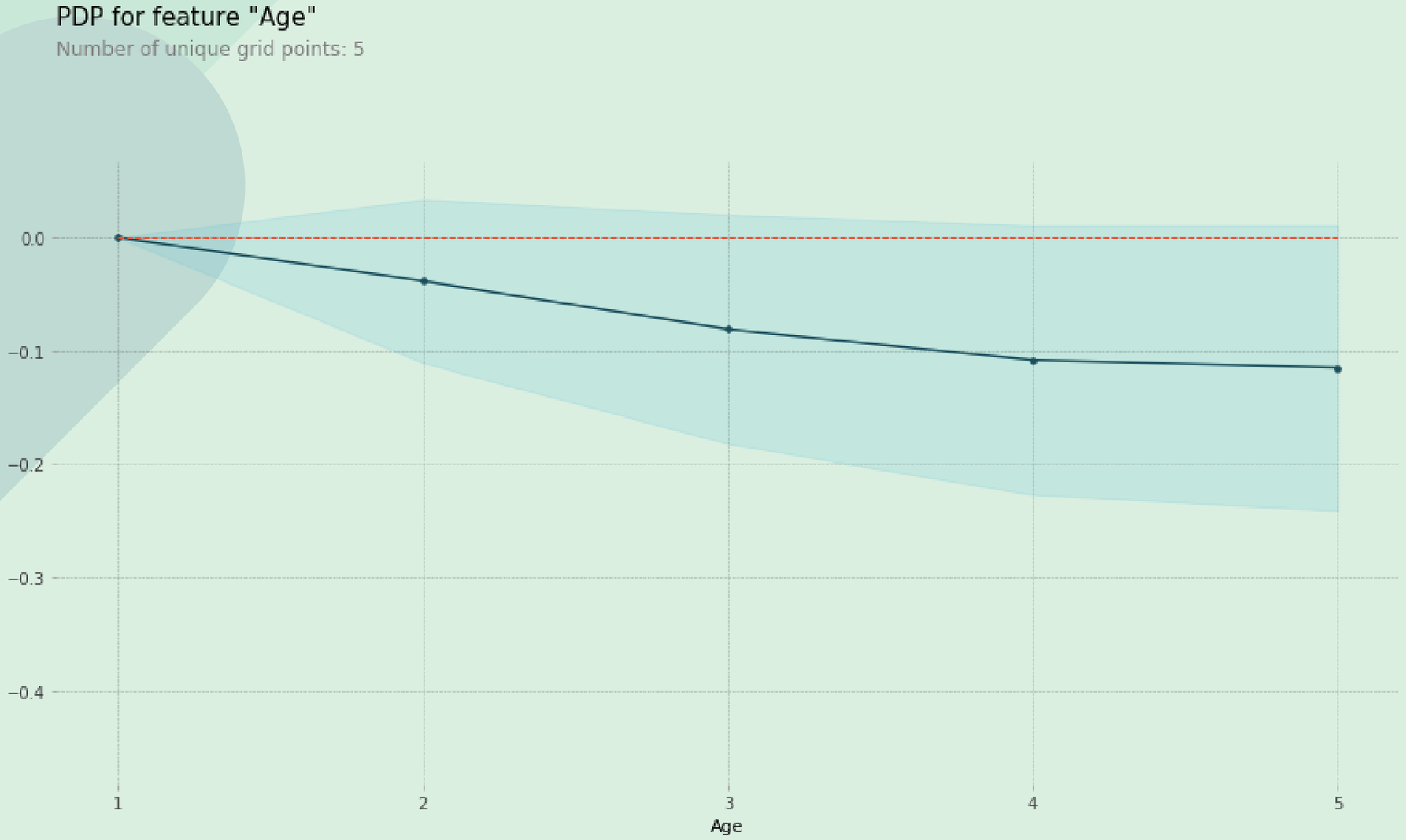
'JoiningYear' PDP 플롯



'Experience' PDP 플롯



'Age' PDP 플롯


PDP for feature "Age"
Number of unique grid points: 5
0.0
−0.1
−0.2
−0.3
−0.4
1
2
3
4
5
Age

7. 결론

가설 검증

1. 최근 입사한 직원일수록 이탈 비율이 높을 것이다.

2. 경험과 이탈 비율은 관계가 없다.

3. 나이가 어린 직원일수록 이탈 비율이 높을 것이다.

결론

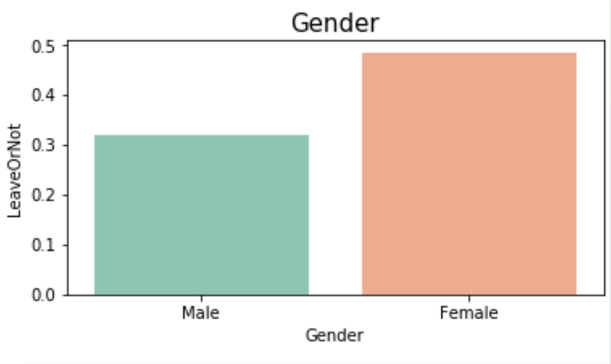
1. 가장 최근인 2018년에 입사한 직원들에 대해 심층적인 설문조사를 진행할 필요가 있습니다.

   > 이들이 퇴사를 결정한 이유를 기반으로 저희 회사를 보완해야 할 것입니다.

2. 비교적 나이가 적은 직원 그룹에서 이탈 비율이 높습니다.

   > 이들이 안정적으로 저희 회사에 정착할 수 있도록 이들의 관심사 파악을 통해 유인이 필요할 것입니다.

결론



추가 결론,

1. 여성의 이탈 비율이 높습니다.

   > 직장 내 부당한 대우를 받고 있지 않는지 등에 대해 조사할 필요가 있습니다.

2. Tier2의 Payment의 이탈 비율이 높습니다.

   > Tier2 그룹의 급여 만족도 등에 대해 조사할 필요가 있습니다.

3. 근무지 Pune의 이탈 비율이 높습니다.

   > Pune 지역 근무지 시설이나 근무지 내 복지 환경을 조사할 필요가 있습니다.

결론

보완 사항

구체적인 임금을 추가하고, 회사, 직무, 상사, 복지 등에 대한 만족도 등을 추가하여 분석하면 보다 의미 있는 분석이 될 것이라고 생각합니다.

하지만, 만족도 설문조사 사용 시, 설문조사의 신뢰도는 생각해보아야 할 것입니다.

감사합니다.