


20230623_Meeting_comment 정리

👤 담당자	 은경 이
☀ 상태	진행 중
📅 마감일	@2023년 6월 23일
■ 태그	Meeting comment

Meeting Comment 정리)

Data EDA 재진행 필요

1) 취득일(ECNY_DT) , 상실일(OUT_DT) 변수 값 형태가 이상한 관측치를 가지는 객체의 기록 전부 제거

(해당 Data Example)

: “ECNY_DT”, “OUT_DT” 변수 길이가 8이 아닌 경우

EWTABLE: Work.Temp						
	연번	성별	취득일	상실일	상실사유	신
	3118948	남	20040711	2005071	폐업,도산,공사중단	Y
	3626498	여	20041022	2004125	계약만료, 공사종료	Y
	1298299	남	19980101	2004021	개인사정으로 인한 자진퇴사	Y
	2220119	여	20020415	2003121	계약만료, 공사종료	Y
	589339	남	19950701	2003071		Y
	1341744	남	20020101	2004101	기타개인사정	N
	1259667	남	20031001	2003101	개인사정으로 인한 자진퇴사	N
	1055350	남	19950701	2005021		Y
	69041	남	20030401	2004030	그밖에회사사정에의한퇴직	N
	2703949	여	19980301	2003121	그밖에회사사정에의한퇴직	Y
	2282795	남	20031104	2004051	기타개인사정	N

2) 재정의한 “DURATION” 변수값이 음수인 관측치를 가지는 객체의 기록 전부 제거

3) 사업장 정보 변수 중 하나라도 결측이 있는 관측치를 가지는 객체의 기록 전부 제거

4) 개인 별 근무일 중첩이 있는 객체(즉, two-job인 객체 의미)의 기록 전부 제거

(Data Example)

168	2800768	남	20170101	20181231		N	75121	75120	75121	019	3	20686153330	2068615333	주식회사 피누스이앤씨	32
169	6915684	남	20170101	20181231	계약만료, 공사종료	N		75120	75121			91918247571	2068615333	(주)피누스이앤씨-강릉2단지한신	27

- 업종, 직종, 직종차수, 우편번호, 종사자수 column이 없는 개인 (또는 row)를 배제하는 게 분석인 편하나, 우선 사업장 정보 변수 missing에 패턴이 있는지 살펴볼 필요는 있음.
- 해당 관측치 전부 추후에는 포함할 필요 있음

5) 추후 보고를 위해 아래 사항 파악할 필요가 있음.

- Raw data의 distinct INDI_ID : xx명
- Data quality check한 후 이상치 존재(사업장 변수 missing도 포함)하는 객체 제외한 후 파악한 distinct INDI_ID : xx명
- 근무기록 중 중첩이 있는 개인 제거한 후 파악한 distinct INDI_ID : xx명
- 최초입사 전 암발병자 제거한 후 파악한 distinct INDI_ID : xx명

Outcome 변수 / 공변량 재정의

1) "개인 암 이력 변수" 재정의 필요

: Outcome이 폐암인 경우, 백혈병도 기타 암 이력에 포함해야 함. (백혈병에 대해서도 마찬가지)

→ Outcome 별 개인 암 이력 변수 따로 생성하는 것이 better

2) "Entry" covariate는 Dummy variable로 재생성

: 즉, "1995~1999년 첫 입사" / "2000 ~ 2004년 첫 입사" ... "2015~2018년 첫 입사" 각각 변수 따로 생성해 1 아니면 0의 값을 가지는 이변량 변수로 재생성하는 것이 better