**2018.08.15 회의록**

**1. 데이터 특성 파악**

1-1 : 데이터한계점

* 추정의 추정
  + 너무 많은 결측치 → 1만개 데이터 추정 → 14만개 데이터 추정
  + 하나의 열 추정 → 이를 토대로 다음 열 추정

1-2 : EDA

* 고객 기본정보를 기준 → 재무정보 분포 탐색
* 연령별, 성별, 소득분위별 등등.. 고객정보 하나당 각 금융정보 변수의 형태 파악
* 기본 정보 중 연관성 있는 열들을 묶어서 탐색 (예) 20대, 미혼의 총자산

2. 방법론 정리(관건 → Data테이블에 없는 유형을 어떻게 추정할것이냐)

2-1. 결측치 추정

* 17000개 테이블(DATA)에 있는 유형을 찾고 각 열값을 평균내어 14만개 테이블(ANSWER) 에 채움(6000개정도) → DATA를 더이상 사용하지 않고 ANSWER에서 분석 후에 채울것인가(분석 및 결측치 처리) : 데이터의 한계점으로 인해 해당 방법론 기각
* 1단계 결측치 추정 방법론
* DATA에서 유형을 찾아서 ANSWER를 채우는 과정을 생략 → Amelia 패키지로 결측치 imputation하기 → 다시 데이터 마사지 해줄것이냐, 말것이냐 → 바로 DATA에서 모델형성(log regression, machine learning) → ANSWER테이블에서 모델을 적용해 바로 predict (DATA에 없는 유형만 predict하고 DATA에 있는 유형은 그대로 가져오기 ..가능하면 예측값과 실제값 비교)
* 결측치 추정목표 : 원데이터에 최대한 가깝게(machine learning, OR 소득분위로 나눠서 regression ) vs 평균치로 예측(log regression)
* EDA가 어떻게 사용될수 있나 ? → feature select할때 도움되도록 EDA를 진행

2-2. 군집 분석을 통한 peer group 형성

* 절대적 거리의 차이를 없애기 위해 모든 열을 정규화 하기 -> 계층적 군집분석(DATA & ANSWER 테이블)
* 군집분석 결과 그래프를 보고 최종 군집 결정
* (선택사항) 판별분석을 통해 peer group간 기준점 찾기
* 완전한 정보(34개열)을 받았을때 특정 peer group이 선택될 수 있도록 만들어져야함.
* 제한된 수의 정보를 받았을 경우 해당되는 여러개의 peer group중 어떤 peer group을 선택할 것인지?
* 14만개 유형을 몇개의 Peer Group으로 나눌까? -> 아주 충분히 그룹간 특성이 반영될 수 있도록 그룹의 개수가 정해져야함.(원론적인 목적)
* -> 계층적 군집분석 OR 딥러닝
* 나머지 단계(Mapping Table, percentile)은 1,2단계 완료시 금방 됨.

1. eda- 승은
2. 결측치 추정- 효영
3. 회귀분석 머신러닝- 승훈
4. 계층적 군집분석 - 종은