# 1년 간의 고객의 데이터를 바탕으로 고객의 잠재성을 파악

1. 전체 고객 중 8분기동안 한 번도 구매가 끊기지 않은 지속고객

이들을 대상으로 피처를 개발함

이유 -> 전체 인원 중 97프로의 인원이며 매출은 98%를 차지하고 있음

2. 지속고객 중 감소고객 정의를 해야하는 것

계절성 제거 후 cagr 계산한 다음(각 분기별로) 회사의 cage보다 작은 지속고객을 감소고객으로 정의했다.

-> 각 기간별 회사의 cagr 값만 알면 됨.

-> 각 기간별 회사의 cagr 변화 그래프

기간 1. 14-1/15-1 학습

기간 2. 14-2/15-2 검증

기간 3. 14-3/15-3 테스트

3. 피처 데이터를 찾기 위한 과정 (우리가 수집해야 할 데이터가 무엇인지 파악)

1. 연령대 : 정확히 알려주시면 감사하겠습니다😊

- 학생 :

- mz세대 : 89-96년 (19세 이하 이런식으로)

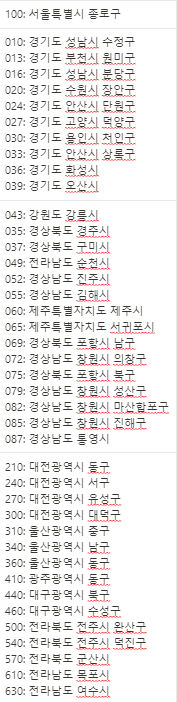
- x세대 : 74-80

- 베이비붐세대: 80년이상

2. 성별

3. 거주지역(우편번호 기준)

: 실제로 데이터 돌릴 때 어떻게 나누셨는지 정확히 알려주시면 감사하겠습니다😊



4. 이용시간: 알려주시면 감사하겠습니다:)

5. 물품 카테고리

- 대한민국 유통물류진흥원에서 제공한 카테고리를 참고해 4개사를 통합함

- 이를 기반으로 순등급(총 4등급을 나눈 후 분기별 차이의 합),

변동율(각 분기마다 변동값의 절대값 합),

비중(의미하는 것이 무엇인지?, 어떻게 구했는지)

6. 분기별? 아니면 전체? <-평균가격에 2배 이상을 고가,

0.5배 이상 2배 이하를 중가,

0.5배 이하를 저가로 지정해

순등급, 변동율, 비중을 구함

7. 카테고리기준 10만원 이상을 **선매품**, 그 이하는 **일상품**으로 해서 순등급, 변동율, 비중 지정 (선매품과 일상품에 대한 나눈 기준을 보여주세요)

**선호물품:** 식품, 디지털가전, 의약품, 기타, 의류를 제외한 사람들의 관심사를 등급 매긴 것

14-1에서 15-1에서 가장 많이 산 것을 기준으로 순등급, 변동율, 비중을 정함

8. 전체 순등급, 변동율, 비중

-> 총 139개의 피처

3-2. 피처 간소화

정확도, 정밀도, 재현율, F1 스코어, ROC AUC스코어를 기반으로 피처 간소화 진행

(진행한 부분에 대한 시각자료 부탁드립니다.)

그래서 34개로 간소화, 컬럼은

'A순등급', '가공식품\_비중', '강원', '경북', '고가\_비중', '고가\_순등급', '구매빈도수증감', '기타\_고가\_순등급', '기타\_중가\_순등급', '대구', '대전', '디지털/가전\_고가\_순등급', '선매품\_비중', '선매품\_순등급', '신선식품\_비중', '신선식품\_순등급', '울산', '의류\_비중', '의약품/의료기기\_비중', '일상용품\_비중', '일상용품\_순등급', '일상품\_비중', '일상품\_순등급', '저가\_비중', '저가\_순등급', '전북', '전체순등급', '중가\_비중', '중가\_순등급', '충남', '학생', '선호물품순등급', '선호물품변동율', '선호물품비중'  
  
3-3. 하이퍼파라미터 -> 그리드서치, 하이퍼오피키(로지스틱, 랜포, xboost, lightgbm) 돌렸음

-> xboost가 가장 괜찮았음

(머신러닝 돌린 시각자료도 부탁드립니다.)

피처 중요도는 전체 순등급, 선매출순등급, 선호물품순등급

증가, 감소한 고객을 나누는 기준(이미 앞에서 cagr기준으로 감소고객을 나눴는데 왜 또 나눴을까요?) 선호물품, 구매 빈도수. 증가를 3개 감소는 4개로 나눴다.

4. 머신러닝 돌리기

14년도 1분기 ~ 15년도 1분기까지 고객의 데이터를 바탕으로 15-2분기 예측 (트레인)

14년도 2분기 ~ 15년도 2분기까지 고객의 데이터를 바탕으로 15-3분기 검증 (벨류)

그 후 테스트