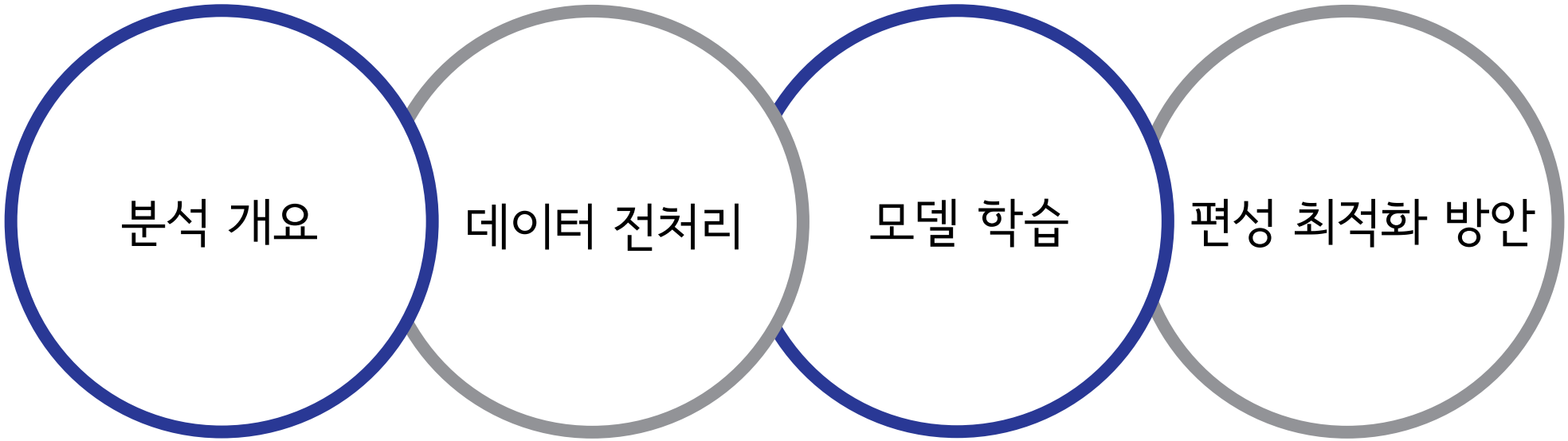
2020 빅콘테스트

데이터분석 분야
챔피언리그

팀명 : 점심 나가서 먹을 것 같아

팀장 - 구교정 (dell0616@gmail.com)
남혜린 (lightsalt28@naver.com)
이 현 (chdnjf103@gmail.com)
정상형 (cookierhkwk@naver.com)

목차



분석 개요

분석 개요

취급액

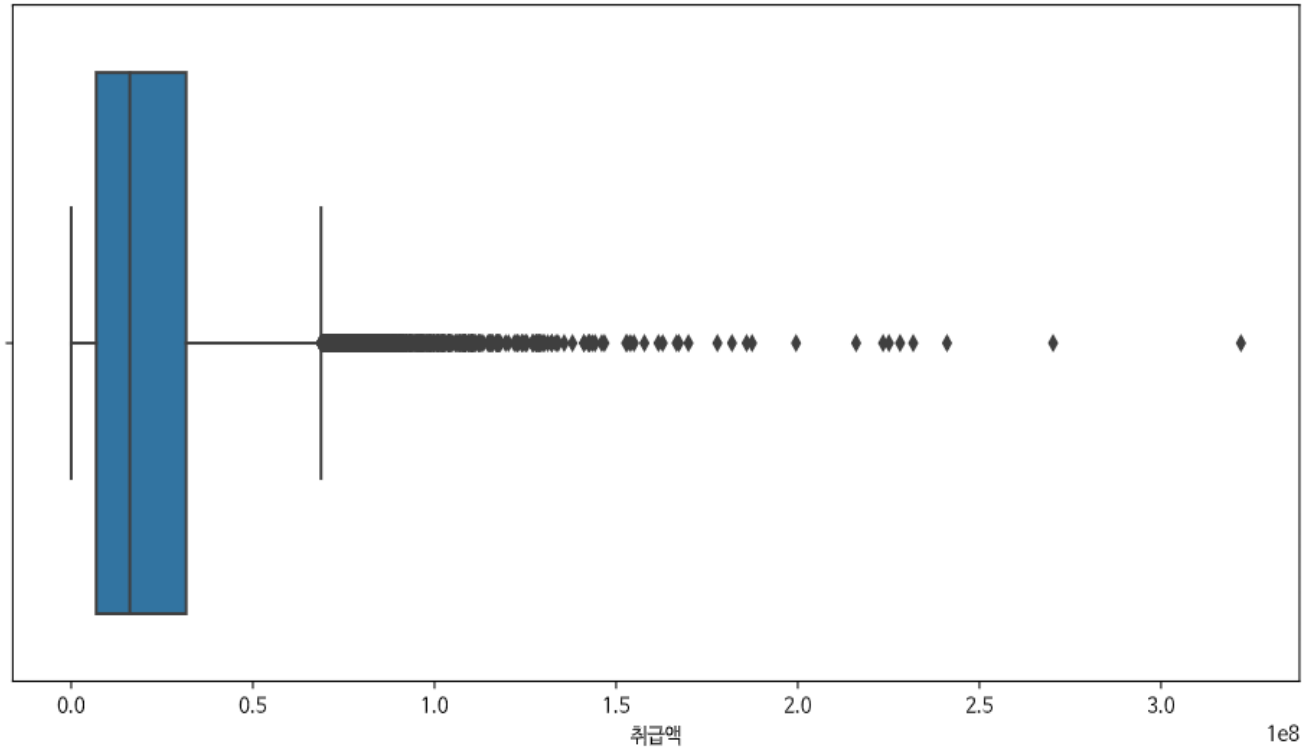
- Target 변수 취급액 분포

  > 매우 넓게 분포

- 단일 모델만으로 예측 무리

  > 데이터 및 모델 분리 필요

취급액 Boxplot



분석 개요

마더코드

마더코드	상품명
100000	엘로엘 아쿠아클린 마스크
100001	국내생산 스텐락 심플 스텐밀폐용기 17종
100002	이보은의 우삽겹 12팩세트

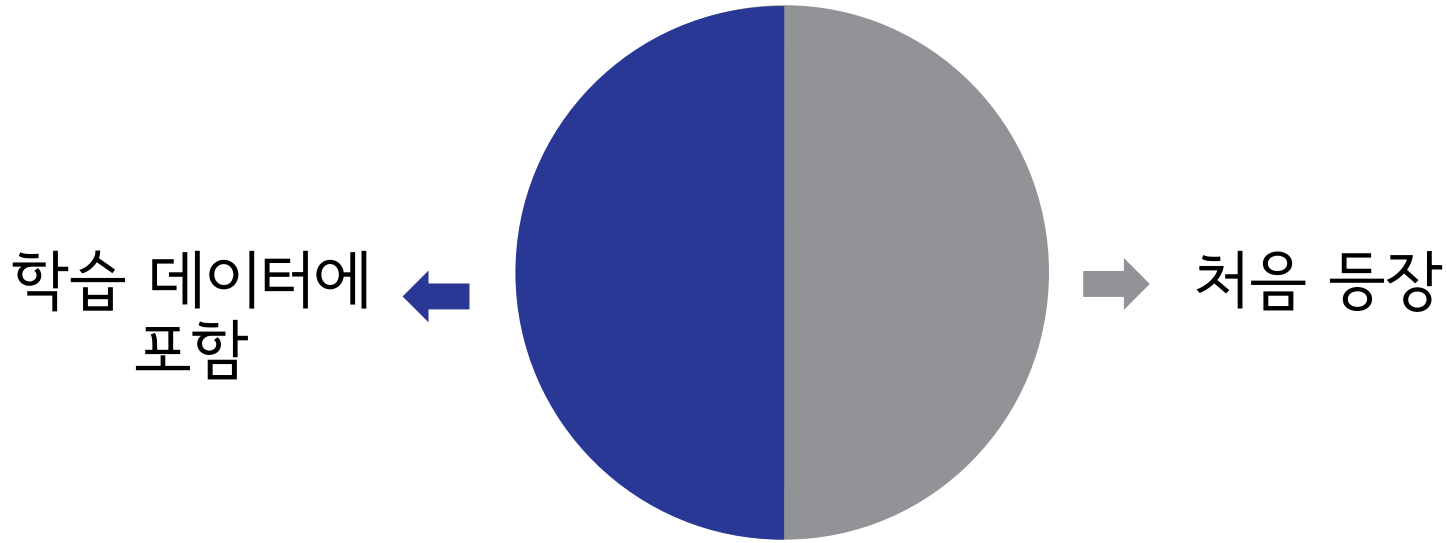
마더코드는 판매하는 상품과 큰 연관이 있는 변수
한번 판매 방송 경력이 있는 상품은 예측 가능성 매우 높음

즉, 예측 정확도 상승에 마더코드는 매우 중요한 변수

분석 개요

마더코드

평가데이터 마더코드



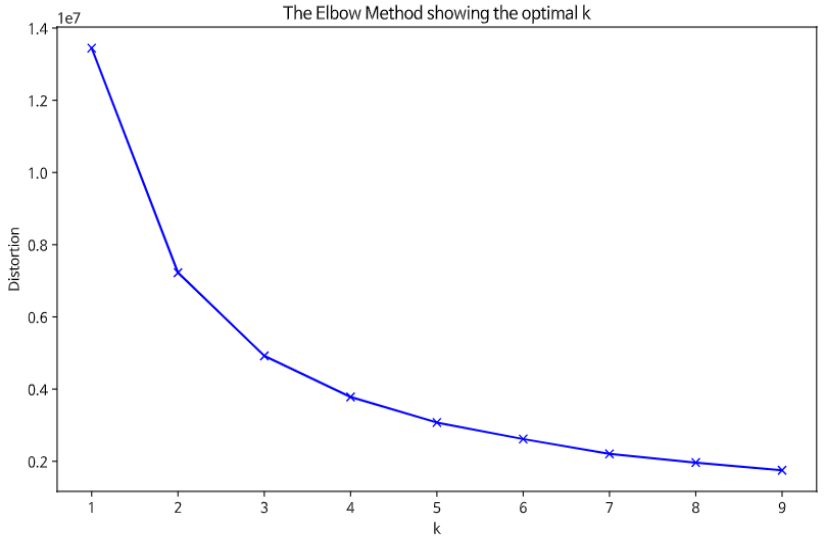
평가데이터의 약 50%의 상품은 처음 판매하는 마더코드

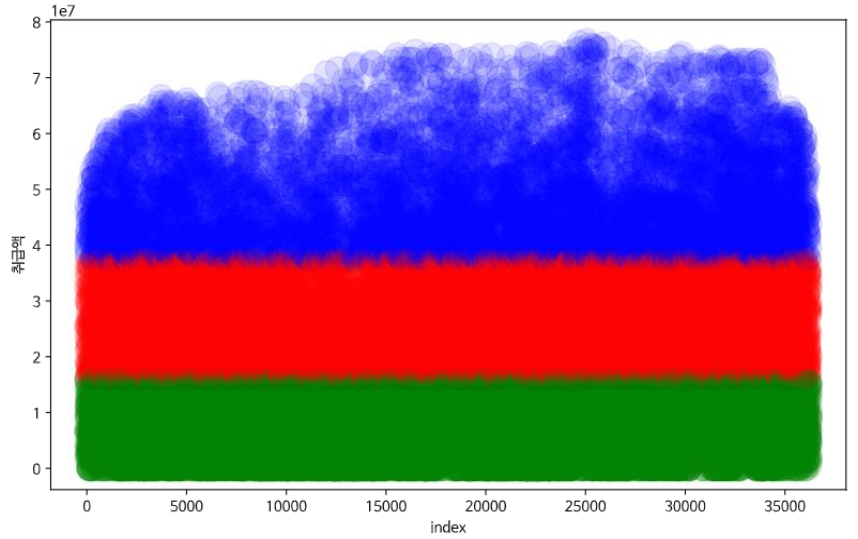
중요한 정보인 마더코드를 활용하기 위한 모델 분리 필요

데이터 전처리

데이터 전처리

군집화





취급액 K-means 클러스터링 학습 결과

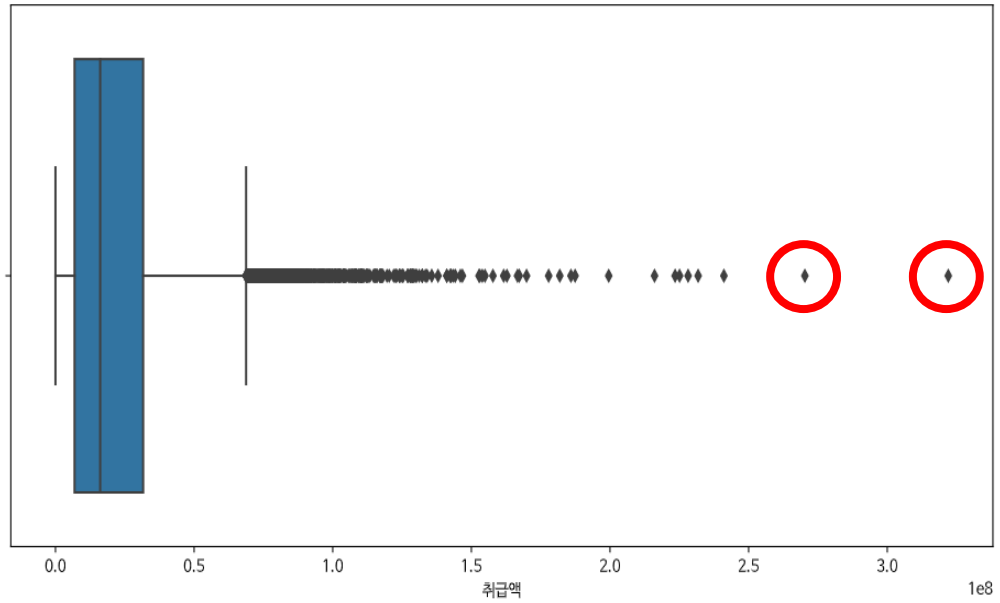
K=3 일 때 가장 적절

예측 군집을 또 하나의 target으로서 사용

데이터 전처리

이상치 제거

학습을 방해하는 이상치 제거



Isolation forest 이용하여

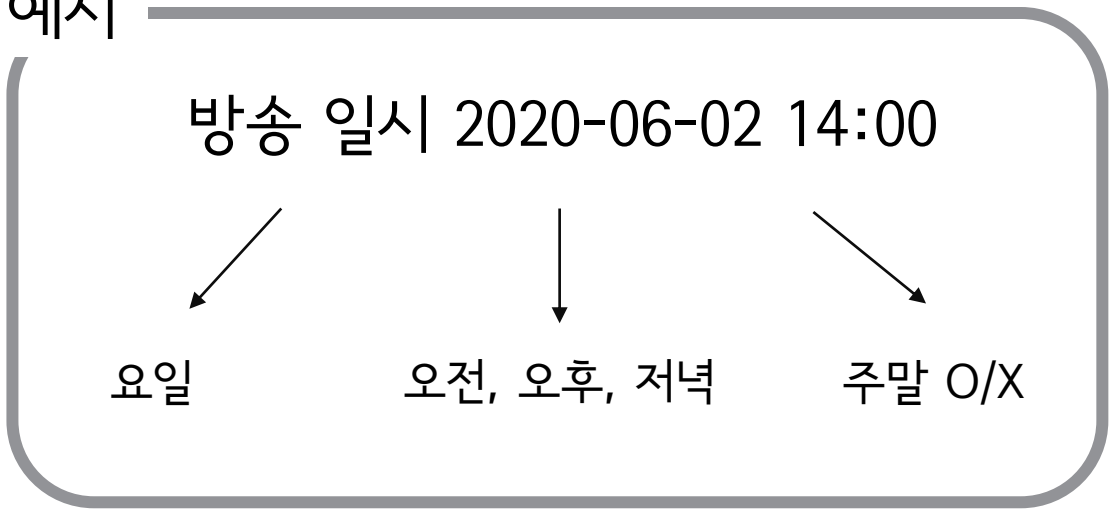
총 3%의 데이터 제거

데이터 전처리

파생 변수

제공 데이터를 이용하여 다양한 파생 변수 생성

예시

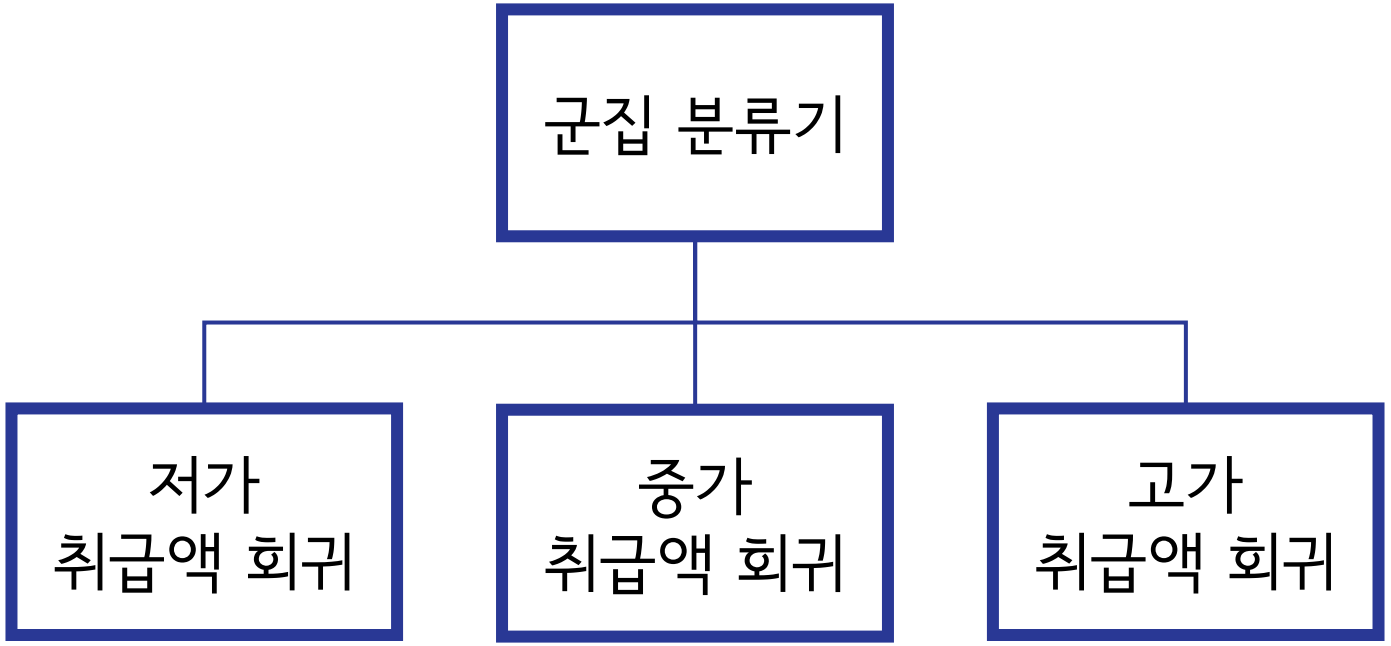


사용한 모든 파생변수는 별도의 '변수설명' 워드파일 참고

모델 학습

모델 학습

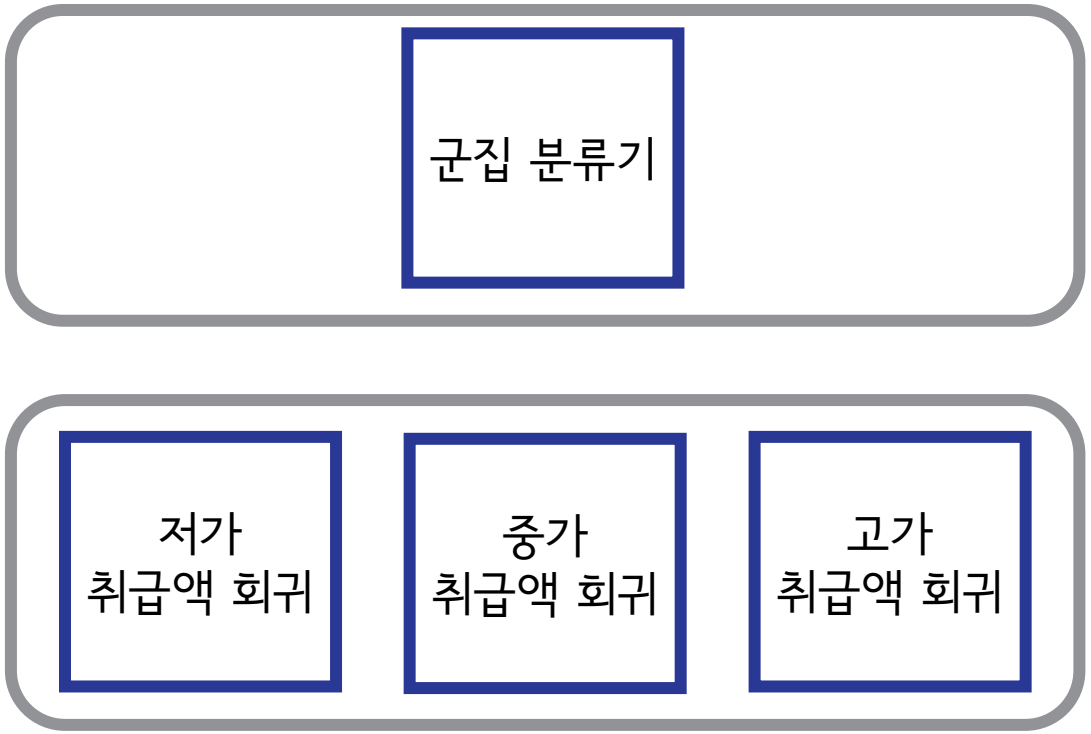
모델 구조



총 4개의 모델로 이루어짐

모델 학습

모델 학습 흐름



군집이 target인 분류기 학습
(Classifier)

각 군집별로 취급액 회귀 모델 학습
(Regressor)

모델 학습

앙상블 학습

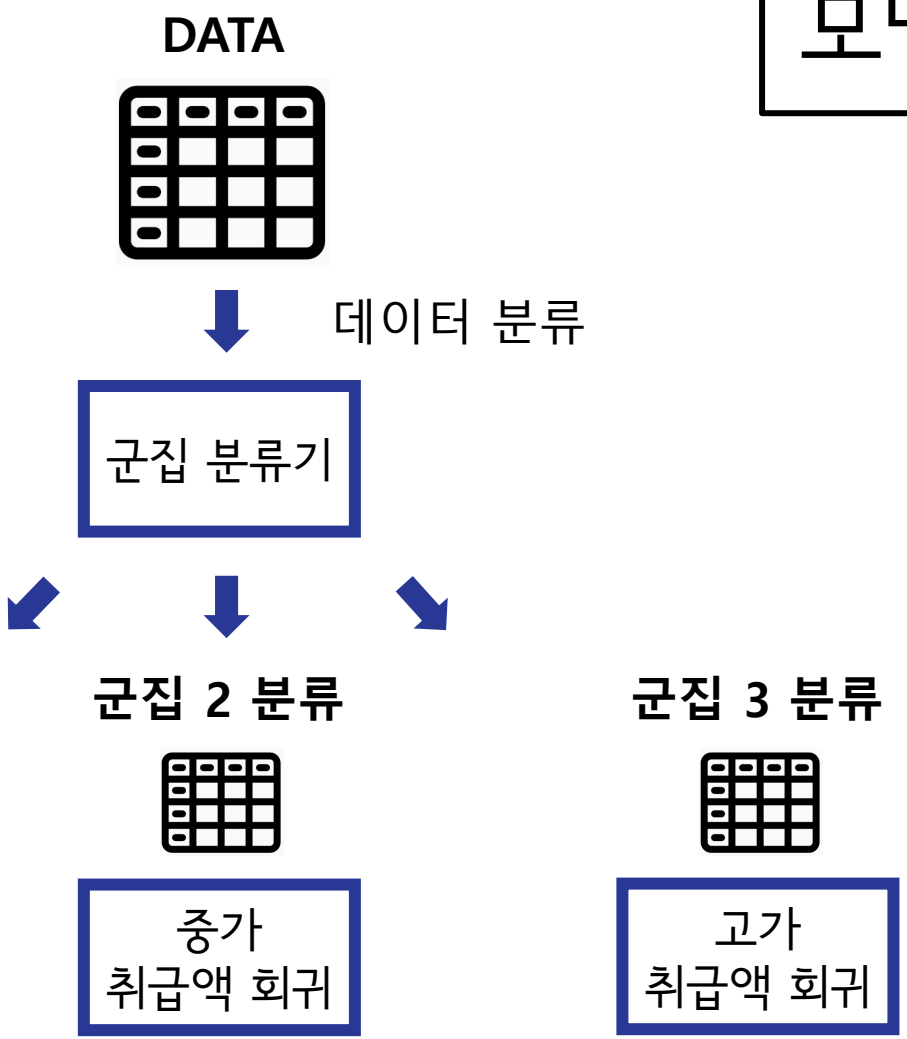
Tree형태의 모델 구조

모델을 세분화와 트리 형태의 모델 구조를 통해

더욱 강건하고 정확한 예측 기대 가능

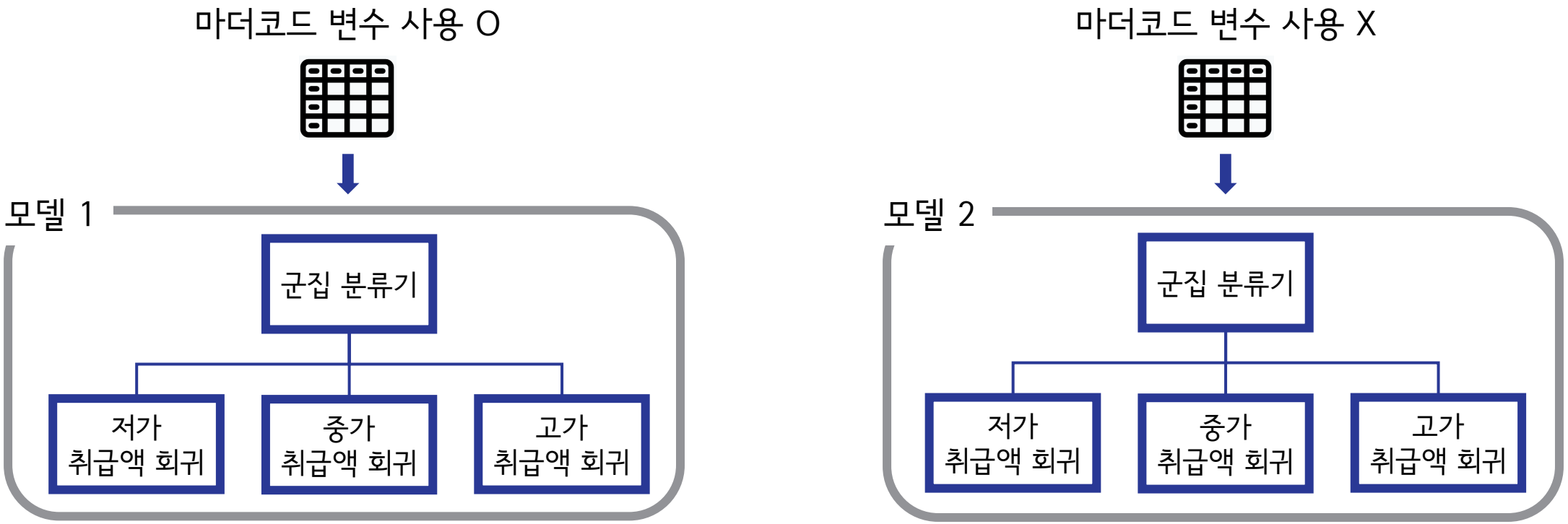
모델 학습

모델 예측 흐름


DATA
데이터 분류
군집 분류기
군집 1 분류
저가
취급액 회귀
군집 2 분류
중가
취급액 회귀
군집 3 분류
고가
취급액 회귀

모델 학습

마더코드 유무



마더코드 관련 변수 사용 여부에 따라
모델을 나눠서 학습 & 예측

모델 학습

마더코드 유무

(예시)

평가데이터의 마더코드가
학습데이터에 등장한 경우

➡

학습데이터로부터 마더코드별
취급액 평균 등의 변수 계산 후 추가

➡

모델 1

평가데이터의 마더코드가
처음 등장한 경우

➡

변수 추가 없음

➡

모델 2

모델 학습

변수 선정

한 변수에 대해 다른 row와
바꿔가면서 예측 결과에 영향을 주는 정도를 파악

데이터 번호	변수 1	변수 2
1	0.5	110
2	0.8	130
3	1.2	80

데이터 번호	변수 1	변수 2
1	0.8	110
2	0.5	130
3	1.2	80

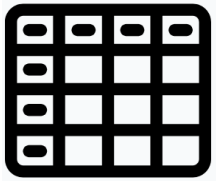
예측 결과 비교 후
변동률이 적은 변수를 중요하지 않다고 판단

이에 따라 변수 선정 및 제거

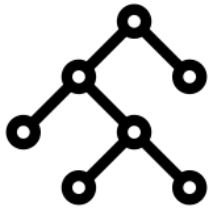
모델 학습

최적 모델 선정

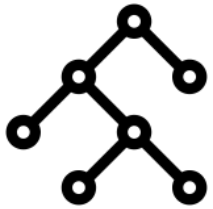
Validation



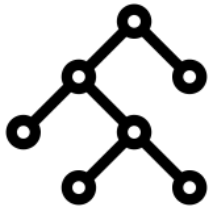
LightGBM



XGBoost



Decision Tree



학습데이터 중 일부를 이용하여
LightGBM, XGBoost, Tree 등 모델들에 대해
최적 파라미터 서치 후 최고 성능 모델 채택

편성 최적화 방안

편성 최적화 방안

기본 전제

방송일시	방송 번호	노출(분)	마더코드	상품코드	상품명	상품군	판매단가
2020-06-01	1	20 x 3	100650	201971	잭필드 남성 반팔셔츠 4종	의류	59,800
	2	20 x 3	100445	202278	쿠미투니카 쿨 레이시 란쥬쉐이퍼&팬티	속옷	69,900
	3	20 x 3	100381	201247	바비리스 퍼펙트 볼륨스타일러	이미용	59,000
	4	20 x 3	100638	201956	램프쿡 자동회전냄비	주방	109,000
	5	20 x 3	100348	201091	벨레즈온 심리스 원피스 4종 패키지	속옷	59,900
	6	20 x 3	100012	200016	AAC 삼채포기김치 10kg	농수축	40,900
	7	20 x 3	100080	200217	아키 라이크라 릴렉스 보정브라 패키지(뉴아키28차)	속옷	99,900
	8	20 x 3	100362	201150	에이유플러스 슈퍼선스틱 1004(최저가)	이미용	39,900
	9	20 x 3	100148	200416	LG 울트라HD TV AI ThinQ(인공지능 씽큐) 55형 55UN7850KNA	가전	1,340,000
			100148	200419	LG 울트라HD TV AI ThinQ(인공지능 씽큐) 65형 65UN7850KNA	가전	1,740,000
			100148	200422	LG 울트라HD TV AI ThinQ(인공지능 씽큐) 75형 75UN7850KNA	가전	2,490,000
	10	20 x 3	100537	201616	[기간]제주바다자연산돔39마리	농수축	39,900

하루에 판매해야 할 [상품] [가격] [노출 시간] 등 각 방송이 정해져 있음을 가정

편성 최적화 방안

편성 최적화

모든 편성 순서에 대해 순서를 바꿔보며 예측함으로써

일일 예상 취급액 평균이 가장 높은 편성 채택

편성 최적화 방안

시뮬레이션

- 다른 시간대 편성에 대한 시뮬레이션 도구로 활용 가능

(예시)

현재 '가전' 상품군은 대부분 저녁 시간에 편성
낮시간에 편성하려면 리스크 감수 필요

학습된 모델을 통해 낮시간으로 변경하여 예측함으로써
새로운 편성 인사이트를 도출할 수 있음

감사합니다