2020 금융 빅데이터 페스티벌

[주제2] 보험금 청구 건 분류

MIRAE ASSET 미래에셋생명

고객 유형화를 통한 보험 청구 적정성 평가 및 추천 서비스 제안

팀 막고라

문성민 김태현 이다은

LOVE AGE

2020 금융 빅데이터 페스티벌

[주제2] 보험금 청구 건 분류

CONTENTS

LOVE AGE

1. 개요

1-1. 공모 주제 탐색
1-2. 분석 방향 설정

MIRAE ASSET 미래에셋생명

보험 청구 ?

청구 적정성(고객 위험도) 판단 후 지급 여부 결정

MIRAE ASSET 미래에셋생명

보험 청구 적정성 판단에 왜 기계학습이 필요할까?

기계학습은 인간이 탐색하지 못하는 패턴을 발견하는데 용이

① 많은 고객 데이터를 복합적으로 보는 것이 어려움

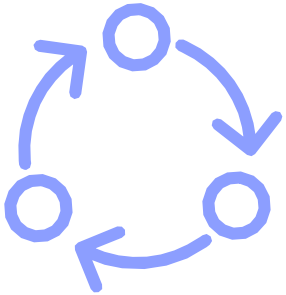
다양한 청구경로 및
다양한 고객정보 등을
하나씩 case화 하여
청구 적정성을 판단하는 것은
많은 Effort 필요

② 정보의 필요 유무에 따라 유연하게 대처 가능

√ 사회적 트렌드

√ 마이데이터 정보를
자유롭게 예측에 활용 가능

Ex) 피쳐 생성과 제거에 유동적



√ 세 Point를 주안점으로 삼음

Pain point

√ 해당 사항들은 Trade-off 관계

√ 한 분야를 강화하면 다른 분야에서
Pain point 발생

Trade off 관계를 고려하며 보험사기 위험 감소에 더 집중하기 위해 Precision을 강조

1. 개요 1-2. 분석 방향 설정(보험사기 위험 감소)

MIRAE ASSET 미래에셋생명

자동지급의 Precision이 낮으면?

자동지급 예측 값 중
실제 자동지급 고객이 적음

↓

이는 장기적으로
선의의 고객의 보험료 상승을 유발

자동지급의 Recall이 낮으면?

자동지급 실제 값 중
자동지급으로 예측되는
고객이 적음

↓

기업의 재심사 인력 투입 증대

1. 개요 1-2. 분석 방향 설정(고객 contribution)

고객 contribution 극대화 : "고객유형화"

기대 효과

√ 특정 고객과 비슷한 타 고객의 정보를 활용 가능

- 군집 내 특성비교를 통해 다양한 효용가치 창출 가능

- 군집을 활용하여 개별 고객마다 집중하는 것 보다 인력 절감 가능

- 보험사기 탐지에 비교군 생성

한계

√ 군집유형분류가 어려운 경우 존재

√ 군집의 중심점과 멀어질수록 왜곡이 발생
(군집이 오히려 고객을 대표하지 못할 수도 있음)

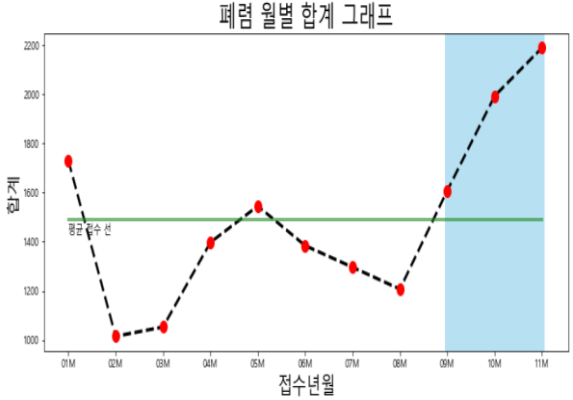
➔ 이를 해결하기 위해 다차원에서 고객을 볼 수 있도록
여러 종류의 군집으로 세분화 필요

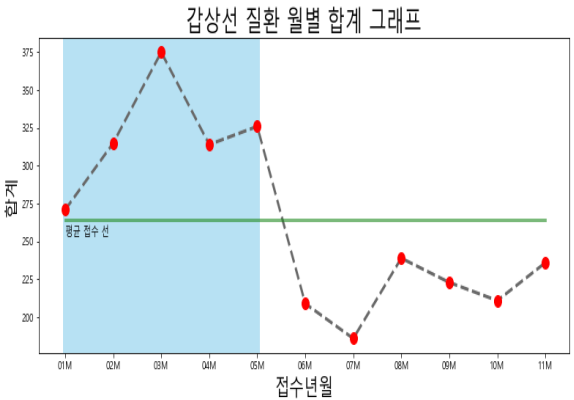
LOVE AGE

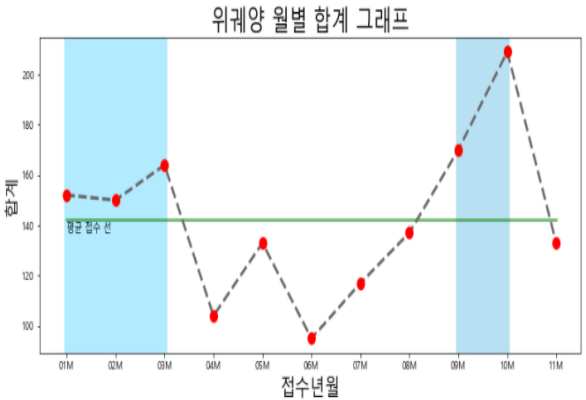
2. 데이터 해석 및 가공

2-1. 질병 트렌드 파악
2-2. 월별 유사성 검정
2-3. EDA 결론 및 한계
2-4. Feature Engineering

2. 데이터 해석 2-1. 질병 트렌드 파악







Ex1) 폐렴

겨울과 환절기에 자주 발생

Ex2) 갑상선 질환

미세먼지의 영향으로
봄에 발병이 증가

Ex3) 위궤양

모두 소화기관 질환으로
통증이 환절기에 심해짐

월별로 질병의 위험률, 발병률이 달라짐
보험금 청구는 이러한 트렌드에 영향을 받을 것임

MIRAE ASSET 미래에셋생명

분산 분석(ANOVA)을 통한 월별 유사성 검정

① 각 질병 별 청구보험금의 평균이 모든 월에 상관없이 동일한가?

① 특정 월의 각 질병별 청구보험금의 평균이 직전 월과 동일한가?

√ 질병별 청구 보험금의 모평균은 월별로 모두 같지 않음

√ 대부분 질병에서 질병별 청구 보험금의 모평균은 직전월과 유사한 모습을 보임

MIRAE ASSET 미래에셋생명

직전 달인 11월만 학습에 사용하는 것이
현 상황에서 질병트렌드를 반영할 수 있는
최적의 선택이라고 판단

But,

√ 주어진 데이터가 11개월로 제한적
→ 질병의 트렌드에 따른 보험 청구량의 변화를 명확히 파악하기에 어려움

√ 예외적인 소수의 질병과 월 데이터가 위 결론을 모두 만족하지는 못했음

∴ 만약 대회가 아닌 현업에서 데이터 사용 기간을 설정한다면,
충분한 데이터를 통해 시계열적 경향성을 재 파악 해야함

MIRAE ASSET 미래에셋생명

청구보험금 : 평균 청구 보험금과의 차이를 통해 비이상적 상황(보험사기 등)을 반영하고자 함
: 향후 군집 이상치 판단에 활용

가입금액 구간별 평균 청구금액 - 청구 보험금

Groupby를 통해 train의 가입금액 구간별 평균 청구 보험금을 계산 후
| 가입금액별 평균 청구 보험금 – 청구 보험금 | 수식 생성, 적용

청구 – 질병 평균 청구액

청구보험금과 해당 질병의 평균적인 청구보험금의 차이를 반영

기타 : 고객 관련 변수 생성

재가입여부

보험료를 납부하지 않고 일정 기일이 경과하면 그 계약은 해지되는데,
이를 다시 회복한다면 보험금을 받고자 부활한 것일 수도 있다고 가정

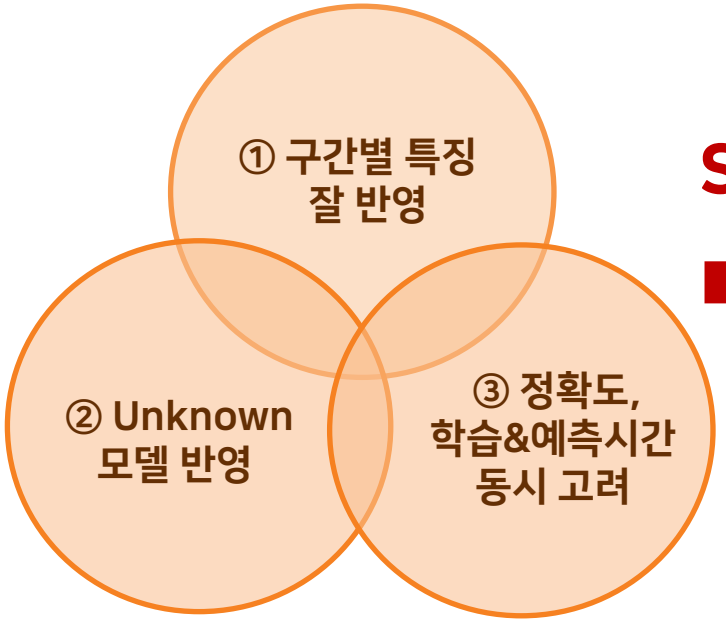
LOVE AGE

3. 모델링

3-1. 모델 계열 선택
3-2. 검증
3-3. 모델 선택 및 고도화
3-4. 모델 해석

MIRAE ASSET 미래에셋생명

모델 계열 선택



Select

Tree 기반의
Ensemble Model 활용

MIRAE ASSET 미래에셋생명

검증 결과

F1 score

1. Extra Tree
2. Random Forest
3. Light GBM

학습 시간

(1 개월 기준)

1. Extra Tree
2. Light GBM
3. Random Forest

예측 시간

(1 개월 기준)

1. Random Forest
2. Light GBM
3. Extra Tree

MIRAE ASSET 미래에셋생명

Extratree안정성보완

√ 모델의 안정성을 높이기 위해 Random state를 다르게 하여 3가지 결과를 반영

새로운 모델 추가

√ 사례기반 모델 추가

- Knn Classifier

2StageModel제안

√ Stage 1 : 심사와 조사를 한 class로 놓고 자동지급과 그 외를 분류

√ Stage 2 : 심사와 조사를 분류

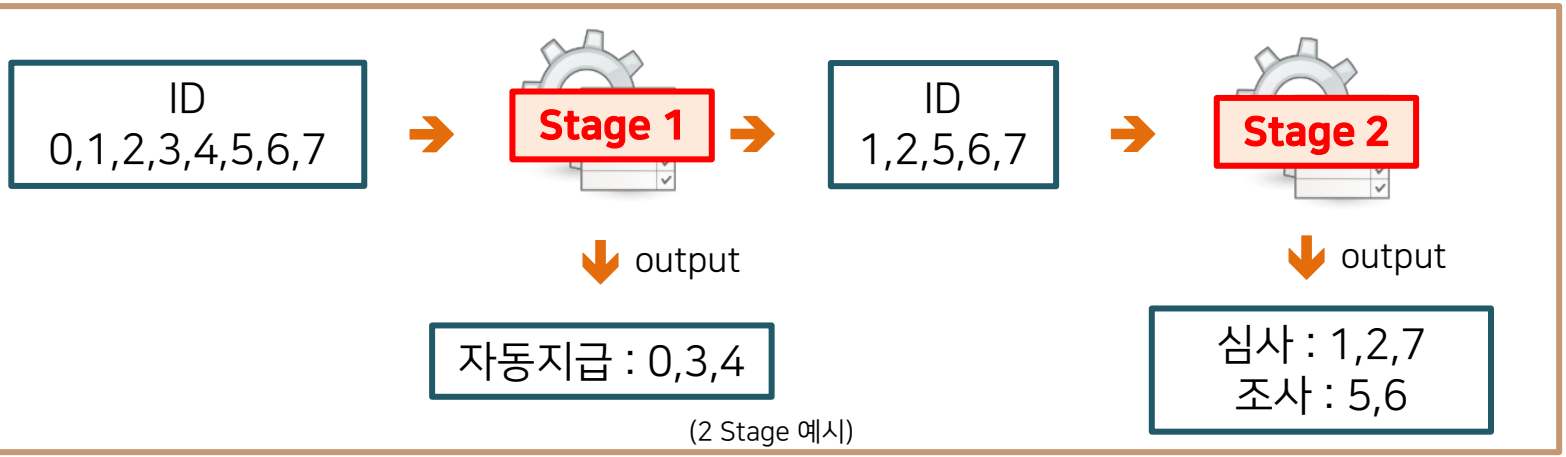
MIRAE ASSET 미래에셋생명

최종 모델 및 평가지표 제안 (2 Stage Model)

√ 심사와 조사를 한 class로 놓고 자동지급과 그 외를 분류하는 이진 분류기를 먼저 구축한 뒤, 그 후 심사와 조사를 분류하는 2Stage model 제안

√ 자동지급에 대한 Precision을 집중하면서 전반적인 성능도 고려

(0.7 * 자동지급의 precision score) + (0.3 * 전체 class의 F1 score 평균)



(2 Stage 예시)

MIRAE ASSET 미래에셋생명

Ensemble 모델의 한계 : 모형이 복잡해 질수록 해석이 어려움

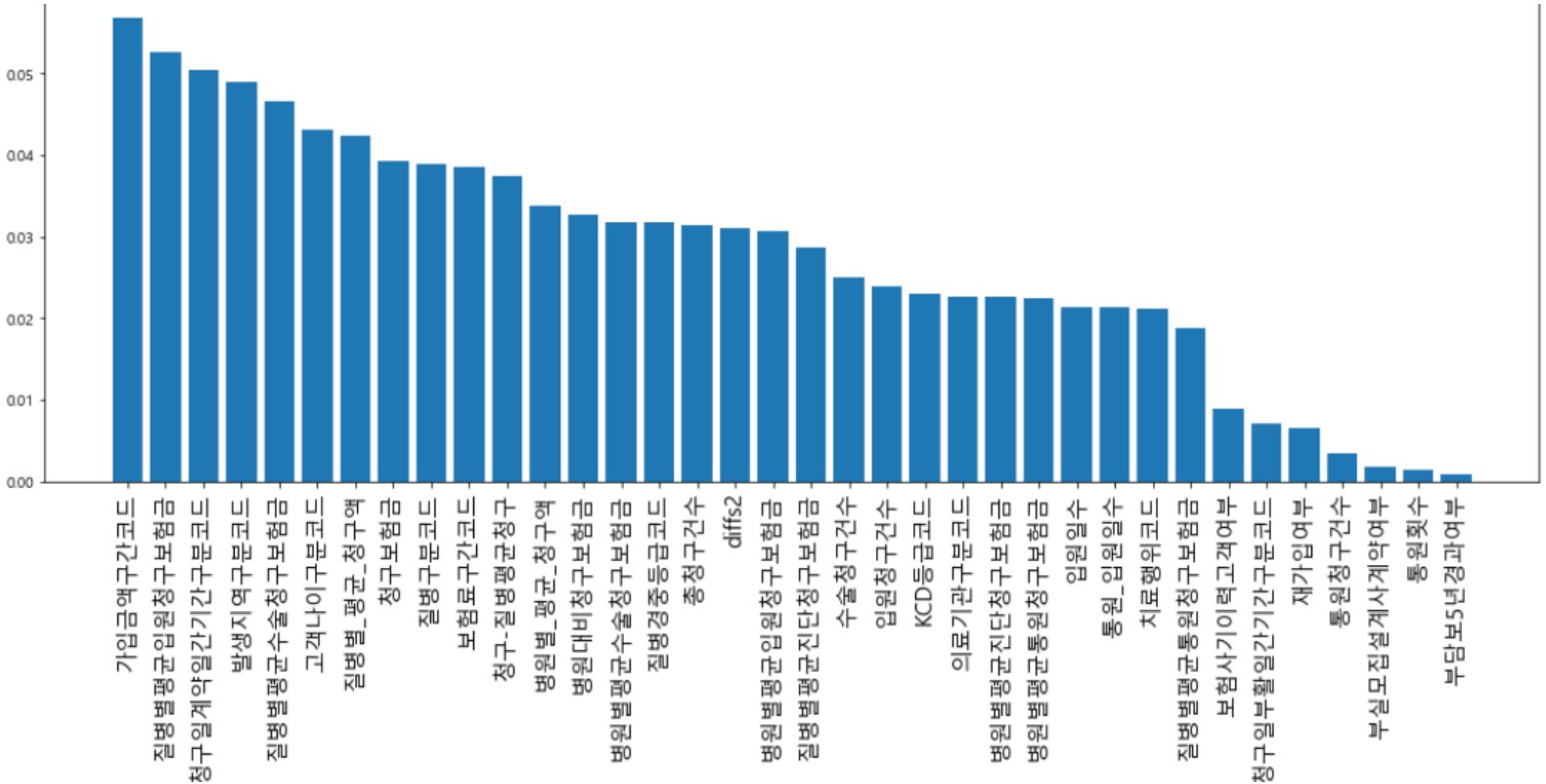
√ 보험사기를 줄이려는 목표에는 부합하나
왜 모델이 보험사기로 예측하였는지를 파악하기 어려움

√ 고객에게 명확한 정보전달이 어려움

[고객 정보 중 Feature Importance가 높은 변수들이 실제로
분류에 어떤 영향력이 있는지 사례분석을 통해 해석]

MIRAE ASSET 미래에셋생명

"Feature Importance"



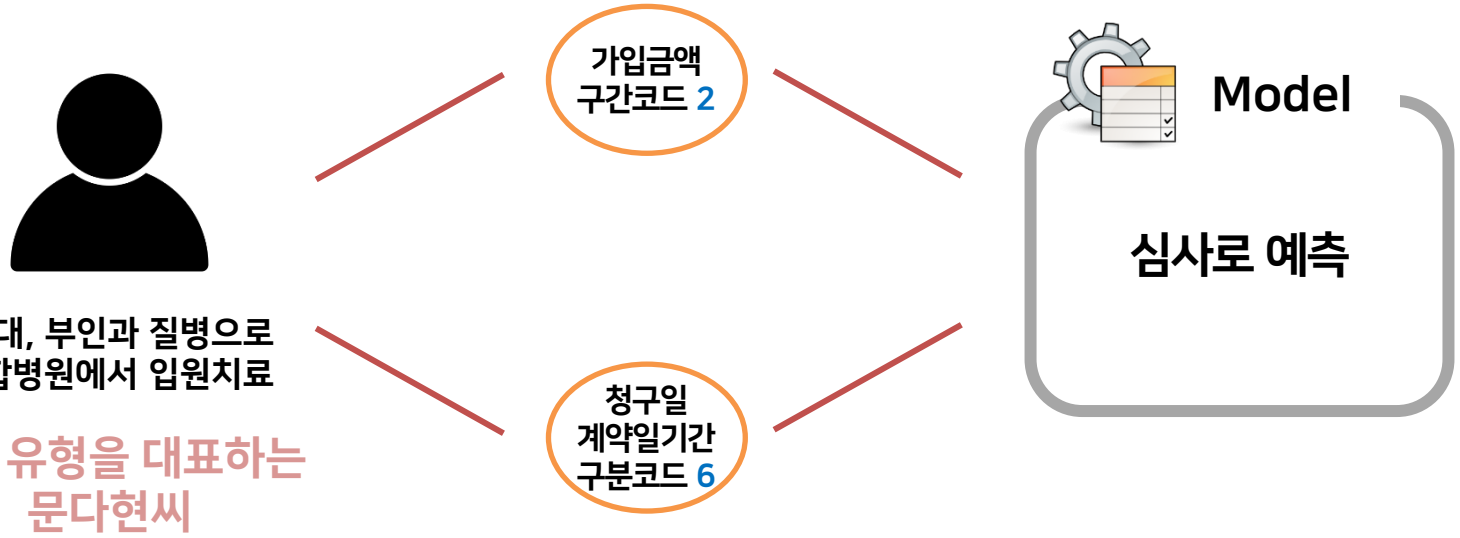
MIRAE ASSET 미래에셋생명

사례 활용 하기

Case

50대인 문다현씨는 부인과 질병으로 불편을 겪어 종합병원에서 입원치료를 진행하였다.
문다현씨의 가입금액구간은 2000만~3000만원 미만, 청구일계약일 기간은 5년 초과로
이 때 2 Stage모델은 문다현씨의 청구에 대해 심사로 예측하였다.

√ Feature Importance가 높은 고객 정보 변수들을 활용



해당 유형을 대표하는
문다현씨

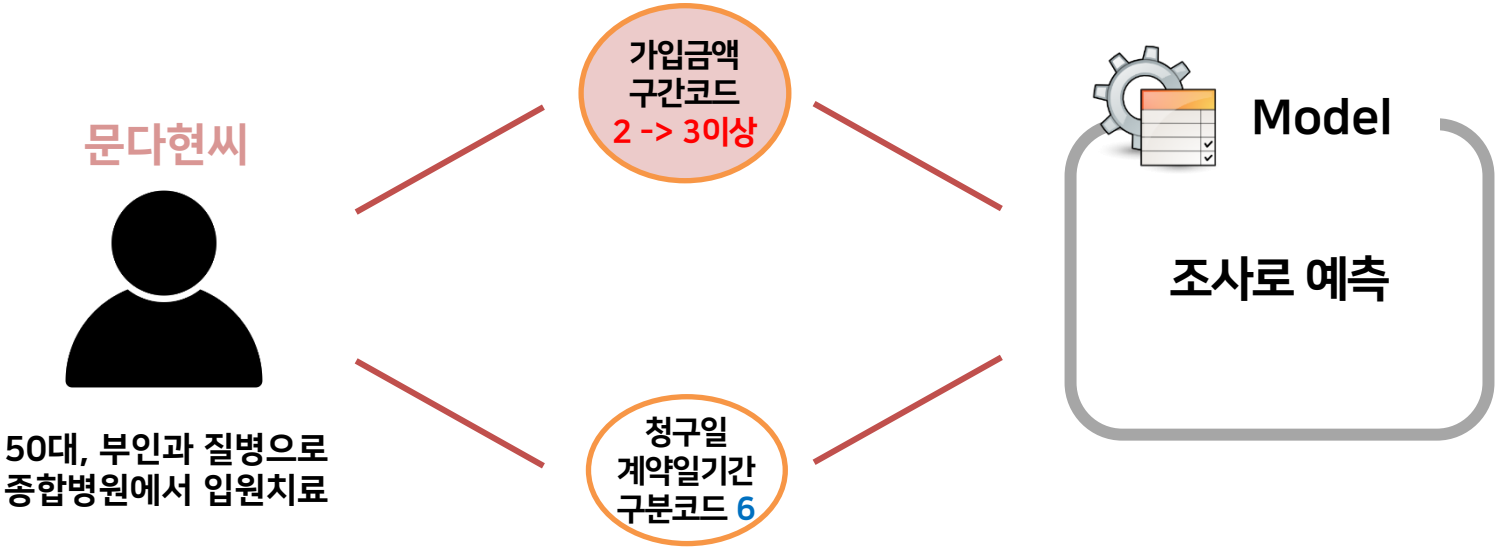
3. 모델링 3-4. 모델 해석(예시)

MIRAE ASSET 미래에셋생명

Case

√ 다른 조건이 동일하다는 가정하에 가입금액이 높아지면, 모델이 심사에서 조사로 예측

√ 즉, 해당 모델에서는 문다현씨 유형의 고객은 가입금액구간코드가 클수록 보험사기에 대해 민감하게 판단

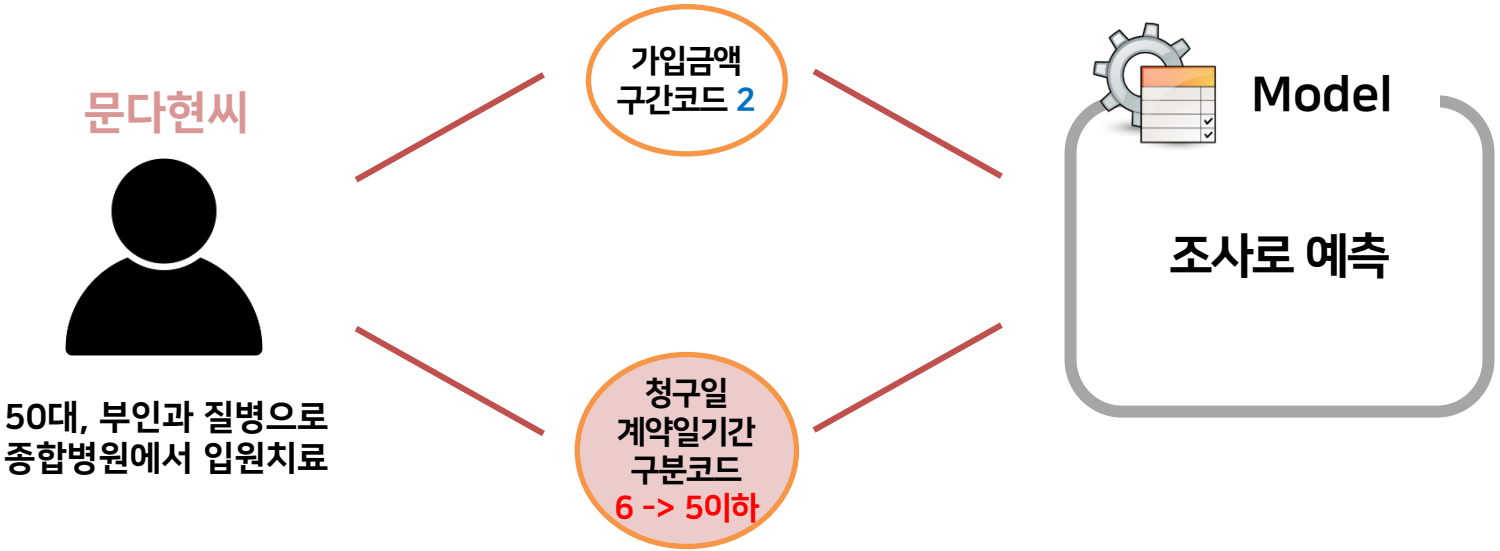


MIRAE ASSET 미래에셋생명

Case

√ 다른 조건이 동일하다는 가정하에 청구일로 부터 계약일 기간이 낮아지면, 모델이 심사에서 조사로 예측

√ 즉, 해당 모델에서는 문다현씨 유형의 고객은 청구일 계약일기간 구분코드가 낮을수록 보험사기의 위험성을 높게 판단



√ 문다현씨와 같은 유형의 고객 또한 위와 같은 해석 적용 가능

MIRAE ASSET 미래에셋생명

기대 효과

√ 중요도 높은 변수를 활용하여 고객 유형별로
청구 적정성 판단에 대한 근거로 사용 가능

√ 고객에게 해당 Class에 분류된 이유를 설명 가능
(ex. 청구일계약일간기간이 낮아 조사로 분류가 되었다고 설명 가능)

√ 모델의 관점에서 객관적인 변수 중요도를 추출하여 현업의 관점에서
중요한 변수를 탐색할 수 있음

LOVE AGE

4. 결론 및 제안

4-1. 최종 모델의 장점 및 한계
4-2. 모델 개선 및 비즈니스 시스템 제안

4. 결론 및 제안 4-1. 최종 모델 장점 및 한계

MIRAE ASSET 미래에셋생명

최종 모델 장점

고객의 유형별 Feature 해석이 가능함

기존의 보험금 청구 심사 프로세스와 가장 유사한 모델

Precision 강화를 통해서 모델의 자동지급 예측 오류를 낮춰
경제적 손실을 방지 및 선량한 고객 피해 최소화

바로 전달의 보험금 청구에 대한 최신 트렌드 반영 가능함

4. 결론 및 제안 4-1. 최종 모델 장점 및 한계

MIRAE ASSET 미래에셋생명

한계

① **비식별화 데이터** : 고객을 유형화 할 수 있는 식별화 정보의 부족으로 인해 다양한 요인의 고객 세분화가 어려움

② **직전 월 데이터 활용** : 직전달에 나타나지 않은 유형에 대한 예측력이 약함

4. 결론 및 제안 4-2. 모델 개선 및 비즈니스 시스템 제안

MIRAE ASSET 미래에셋생명

고객 군집화 시스템 구축 제안

1. 배경

√ 식별화 정보 추가로 기존 모델 개선

√ 앞으로의 보험사기는 언택트(Untacted) 가속화 추세에 따라 디지털 환경 중심으로 더욱 확대될 것으로 예상됨. 이에 대응할 수 있는 기계학습 시스템 구축 필요성 제기

√ 고객 맞춤형 서비스의 필요성 제기

2. 목적

기존 모델의 개선 + 효율적인 보험 사기 예측 지표 생성 + 개별화된 고객 서비스 지원

3. 내용

다양한 식별화 데이터를 넣은 클러스터링 기법을 활용해 고객 군집을 생성
이를 활용한 사후 분석 및 서비스 제공

4. 결론 및 제안 4-2. 모델 개선 및 비즈니스 시스템 제안

MIRAE ASSET 미래에셋생명

4. 프로세스 식별화 데이터를 활용해 고객 군집을 생성

식별화 된 고객 데이터

보험 가입 시 입력하는 고객 데이터

+

mydata

고객의 의료정보, 금융정보 등 자유롭게 수집 가능 전제

고객 클러스터링 모델 구축

1. 소비성향 군집
2. 건강정보 군집
3. 질병관리 군집

[결과]

고객의 군집 정보를 우리 모델 데이터의 한 특성변수로 삽입 → 기존 모델의 성능 개선 및 안정화

군집별 사후분석 + 이상치 탐지 → 보험사기 예측의 지표로서 활용 (신호등 지표)

유사 고객 군집 정보와 기존 온라인 서비스 연계 → 군집을 기반으로 커스터마이즈 된 보험상품의 추천

4. 결론 및 제안 4-2. 모델 개선 및 비즈니스 시스템 제안 (클러스터링)

MIRAE ASSET 미래에셋생명

고객 클러스터링 구축

기존에 보유한 정보(나이, 지역 등)를 기반으로
마이데이터를 활용하여 세 가지 군집 구축

군집 1
"질병 관리 군집"

과거 질병, 치료행위

건강검진 정보

웨어러블 기기를 통한 생활패턴 수집

의료기관 제휴, 헬스케어 기업 제휴

군집 2
"자산 정보 군집"

생활수준, 자산현황

수익, 예적금 내역
(자산관리 성향)

금융사 제휴

군집 3
"소비 성향 군집"

구매이력

카드사 제휴

4. 결론 및 제안 4-2. 모델 개선 및 비즈니스 시스템 제안 (기존 모델 보완)

MIRAE ASSET 미래에셋생명

기존 Data set

	나이구분코드	지역구분코드	가입금액구간	…
1	4	1	3	…
2	2	4	6	
…	…	…	…	

+

군집 정보 추가

	질병군집	자산군집	소비생활	
1	11	3	8	…
2	2	5	9	
…	…	…	…	

↓

기존 모델의 성능 개선 및 안정화

4. 결론 및 제안 4-2. 모델 개선 및 비즈니스 시스템 제안 (위험도 탐지 신호등)

MIRAE ASSET 미래에셋생명

위험도 탐지 신호등 이상치 판단 신호등

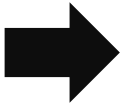
√ 최종 모델의 두 과정에서 적용 됨

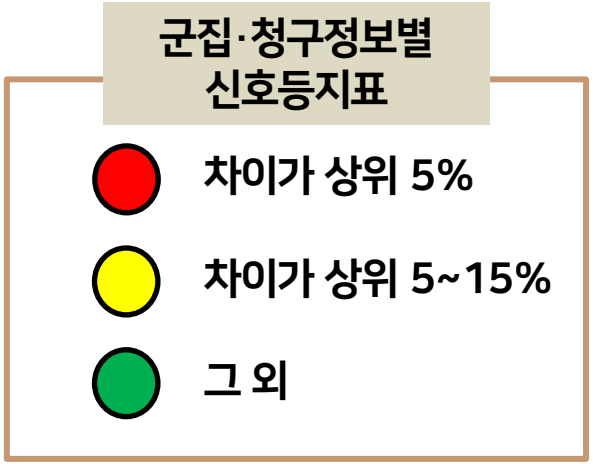
Stage 1) 자동지급 예측에 대해 고위험군을 재검토하여 자동지급 혹은 심사 결정

Stage 2) 조사 예측에 대해 고위험군을 재검토하여 일반조사 혹은 특수조사 결정

군집	질병구분코드	청구보험금 평균
1	1	3.452
	2	1.683
	3	0.234
	…	…

| Aggregation 평균 – 개별 고객 청구보험금 |





4. 결론 및 제안 4-2. 모델 개선 및 비즈니스 시스템 제안 (위험도 탐지 신호등)

MIRAE ASSET 미래에셋생명

위험도 탐지 신호등

고객 A씨

Stage 1

output

자동지급으로 분류

(자동지급이 아닐경우)

Stage 2

output

심사 또는 조사로 분류

군집 1 군집 2 군집 3

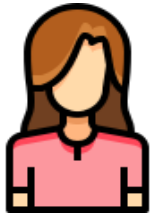
높은 위험군으로 분류

재검토를 통해 자동지급 유지 혹은 심사로 분류

4. 결론 및 제안 4-2. 모델 개선 및 비즈니스 시스템 제안 (위험도 탐지 신호등)

MIRAE ASSET 미래에셋생명

위험도 탐지 신호등



고객 A씨

Stage 1

output

자동지급으로 분류

(자동지급이 아닐경우)

Stage 2

output

심사 또는 조사로 분류

군집 1 군집 2 군집 3

+

낮은 위험군으로 분류

자동지급 유지

4. 결론 및 제안 4-2. 모델 개선 및 비즈니스 시스템 제안 (위험도 탐지 신호등)

MIRAE ASSET 미래에셋생명

위험도 탐지 신호등

고객 B씨

Stage 1

Predict = 1

Stage 2

output

조사로 분류

+

군집 1 군집 2 군집 3

높은 위험군으로 분류

특수조사로 분류

4. 결론 및 제안 4-2. 모델 개선 및 비즈니스 시스템 제안 (위험도 탐지 신호등)

MIRAE ASSET 미래에셋생명

위험도 탐지 신호등

고객 B씨

Stage 1

Predict = 1

Stage 2

output

조사로 분류

+

군집 1 군집 2 군집 3

낮은 위험군으로 분류

일반 조사로 분류

4. 결론 및 제안 4-2. 모델 개선 및 비즈니스 시스템 제안 (보험 추천 시스템)

MIRAE ASSET 미래에셋생명

1) 질병 정보 군집

군집 내 질병 유사도 기반
위험 질병 예측

2) 자산 정보 군집

자산 현황과 투자 성향 분석

3) 소비 성향 군집

소비 패턴 분석으로
노후 필요 자금 예측

보장 질병?

보험 상품?

필요 자금?

군집 기반 맞춤형 보험 추천 가능

4. 결론 및 제안 4-2. 모델 개선 및 비즈니스 시스템 제안 (보험 추천 시스템)

MIRAE ASSET 미래에셋생명

1) 질병 정보 군집 활용 군집내 고객의 질병 유사도 기반 질병 예측

➡ Item Based Collaborative Filtering(아이템 기반 협업 필터링) 사용

① 군집 내 User – 질병 Sparse Matrix 생성

해당 고객이 보험 청구했던 질병

같은 질병 군집에 속한 고객 ID

	췌장암	유방암	간암	대장암	
1	1	0	0	0	…
2	0	0	0	1	
…	…	…	…	…	

➡ Transpose 후 행렬 곱

② Cosine Similarity 기반 군집 별 질병 유사도 행렬 생성

	췌장암	유방암	간암	
췌장암	1	0.693	0.453	…
유방암	0.693	1	0.342	
간암	0.453	0.342	1	
…	…	…	…	

4. 결론 및 제안 4-2. 모델 개선 및 비즈니스 시스템 제안 (보험 추천 시스템)

MIRAE ASSET 미래에셋생명

③ 유사도 행렬 기반 고객의 질병 예측

	췌장암	유방암	간암	…
청구 여부	0	1	0	…

문다현씨 청구 이력 데이터
(개별 고객 데이터)

	췌장암	유방암	간암	
췌장암	1	0.693	0.453	…
유방암	0.693	1	0.342	
간암	0.453	0.342	1	
…	…	…	…	

문다현씨의 해당 군집내 질병 유사도 행렬

질병 유사도 곱
높은 순으로
정렬 후
TOP n 질병 선정

질병 예측

갑상선암

여성 생식기암

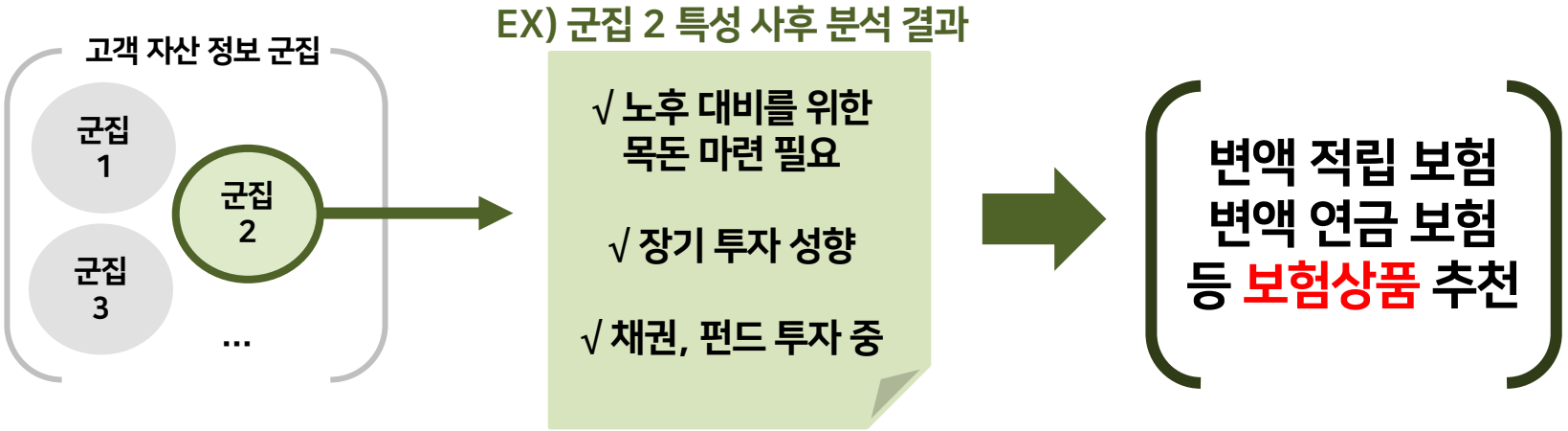
같은 군집 내 고객의
질병정보를 활용해
문다현씨의 위험 질병 예측

4. 결론 및 제안 4-2. 모델 개선 및 비즈니스 시스템 제안 (보험 추천 시스템)

MIRAE ASSET 미래에셋생명

2) 자산 정보 군집 활용

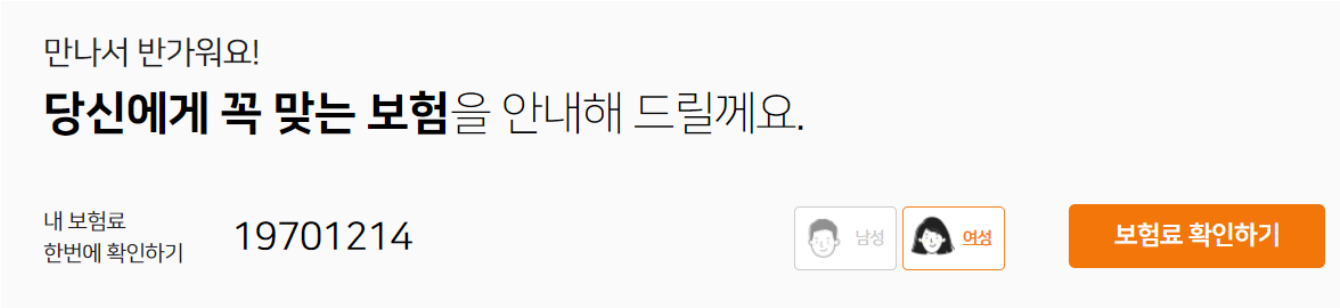
군집별 고객의 **자산정보**와 **투자성향** 분석 후
맞춤형 보험상품 추천 + 온라인 서비스 연계



4. 결론 및 제안 4-2. 모델 개선 및 비즈니스 시스템 제안 (보험 추천 시스템)

MIRAE ASSET 미래에셋생명

실제 서비스 적용도 가능



기존 미래에셋 온라인 보험 추천시스템에 투자성향 관련 간단 문항 추가

↓

수집한 투자성향 특성들을 바탕으로
기존 자산 정보 군집 중 하나에 실시간 매핑

↓

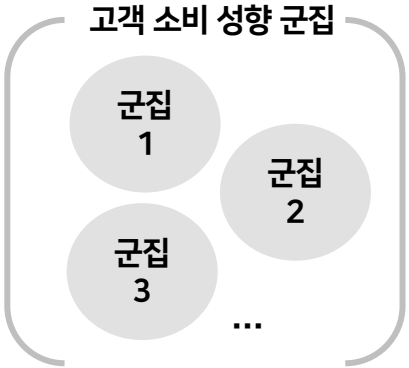
군집에 해당하는 보험상품, 특약 세분화 추천

4. 결론 및 제안 4-2. 모델 개선 및 비즈니스 시스템 제안 (보험 추천 시스템)

MIRAE ASSET 미래에셋생명

3) 소비 성향 군집 활용

군집별 고객의 소비패턴 분석으로
노후 필요자금 예측 + 노후 설계 제안



월별 카드 사용 정보
→ 월별 소비 패턴 분석 가능

이대로 앞으로 계속 소비한다고 가정할 때,
노후 자금이 얼마나 필요한지 예측 가능

(물가 상승률 감안)

필요 노후 자금에 따른 연금 보험 추천과
노후 설계 제안

기대 효과

고객

- 위험질병, 투자성향, 소비성향 등에 따른 맞춤형 보험 설계 서비스 수혜

→ 효율적인 의사결정 도움
고객 만족도 증대

회사

- 고객 만족도 증가로 고객 유지율(PPR) 상승 효과

- 세분화된 맞춤형 추천을 통해 마케팅 효과 증대

5. 부록

5. 부록

MIRAE ASSET 미래에셋생명

참고문헌

< 논문 >
- Extremely randomized trees(Pierre Geurts, 2006)
- 계층화 분석기법을 이용한 건강보험 부당청구 감지지표 우선순위 도출(박민규, 2020)

< URL >
- 미래에셋 Q&A (https://life.miraeasset.com/home/index.do#MO-HO-030402-010000)
- 금융감독원 http://insucop.fss.or.kr/fss/insucop/define02.jsp

< 기사>
- [중앙일보] 진료영수증 학습한 AI가 보험금 심사…한화생명, 기술특허 획득(안효성, 20.09.21)
- [대한데일리] 매년 느는 보험사기, 보험사 AI로 대응(임성민, 20.09.10)
- [청년일보] "보험사기 AI로 잡는다"…KB손보, AI 보험사기 탐지시스템 개발(강정욱, 2020.10.28)
- [연합인포맥스] 보험사기 예방 나선 미래에셋생명, 도수치료 청구율 76% 감소

사용 라이브러리 버전

- Pandas 0.25.1, Numpy 1.16.5, Matplotlib 3.1.1
- Seaborn 0.11.0, Sklearn 0.23.1

MIRAE ASSET
미래에셋생명

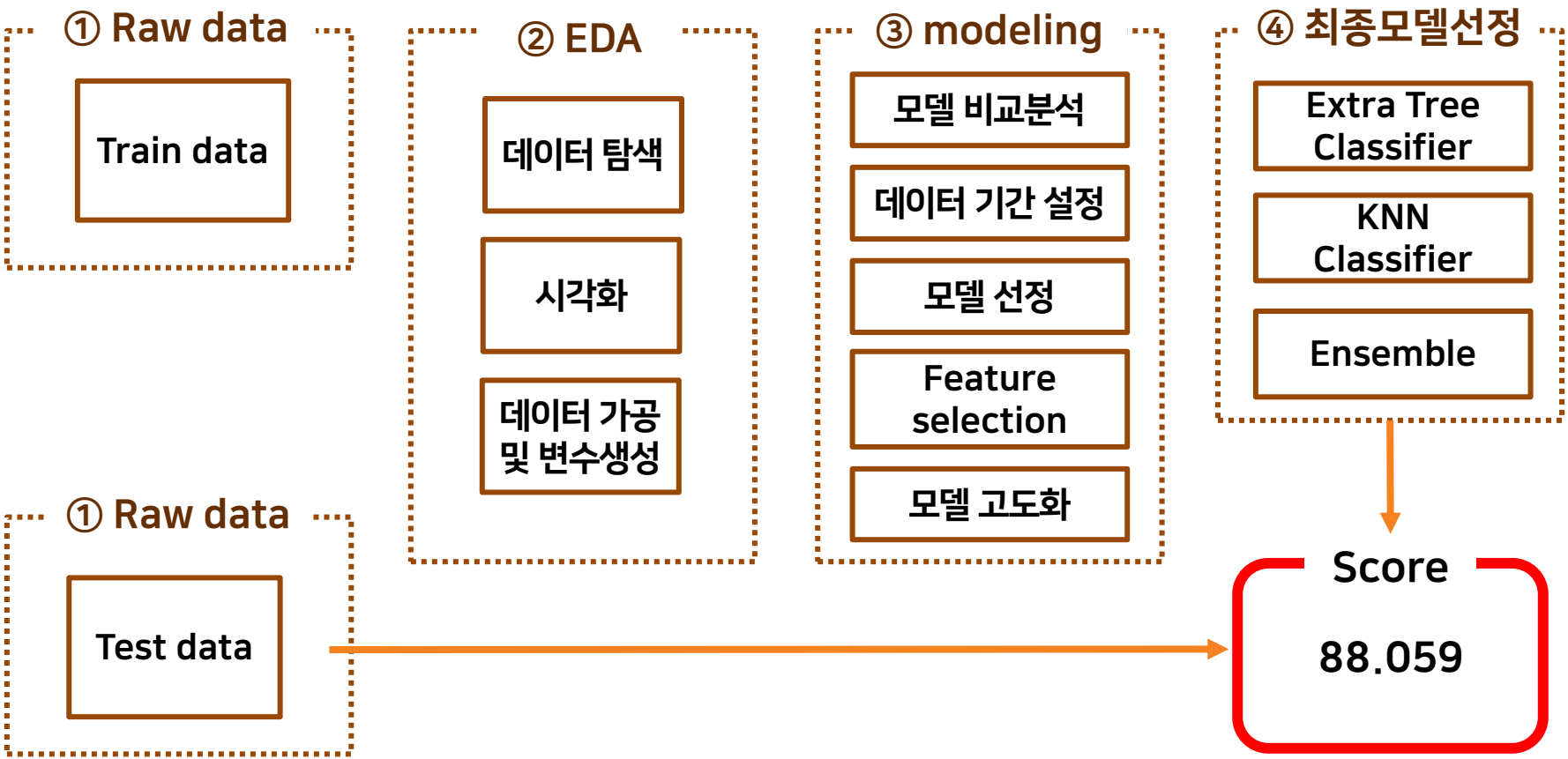
감사합니다

LOVE AGE

Appendix CONTENTS

MIRAE ASSET 미래에셋생명

전체 분석 프로세스



2. 데이터 해석 및 가공 2-1-1. 전반적 탐색

MIRAE ASSET 미래에셋생명

전반적 특징 고려

√ Target의 분포가
비교적 불균형
+ 특성변수간 상관관계 존재

√ 결측 대신 unknown 값
9, 99형태로 표시

√ Train data가 37만개로
충분한 상황

EDA 주안점 도출

√ Target별로
각 특성변수들의 분포
+ 세부적 Heatmap 파악

√ unknown만의
분포 파악 후, 처리 방안 고안

√ Train data의 경향성
파악 후 데이터 정제
방안 고안

이들은 모델 선정의 근거로 작용

2. 데이터 해석 및 가공 2-1-2. 전반적 탐색

MIRAE ASSET 미래에셋생명

HOME > 생활/건강 > 건강

민감한 갑상선, 봄 황사와 미세먼지 배출하는 갑상선에 좋은 음식은?

윤정원 기자 | 승인 2018.05.21 10:23 | 댓글 2

출처: 베이비 뉴스

십이지장궤양의 특징적인 증상은 공복 시 타는 듯한 심와부의 동통 또는 불편감이며, 환자의 60~80%에서 발생합니다. 산이 계속 분비되는 식후 2-3시간 후, 또는 산의 분비가 제일 많은 밤 11시에서 새벽 2시 사이의 야간에 증상이 심해져서 잠을 깨는 경우가 많고, 음식을 먹거나 제산제를 복용하면 증상이 쉽게 사라지는 특징을 갖고 있는데, 매우 심한 증상은 대개의 경우 일시적이며, 병발되는 복부불편감이 30분에서 2시간 정도 지속됩니다. 통증은 만성적이고 주기적(환절기 특히 봄·가을)으로 발생하며 쑤시는 듯, 타는 듯, 물어뜯는 듯 또는 칼로 베는 듯한 느낌이 있습니다.

출처: 삼성 서울병원

2018년 폐렴 환자 수 134만 명, 환절기에 늘어

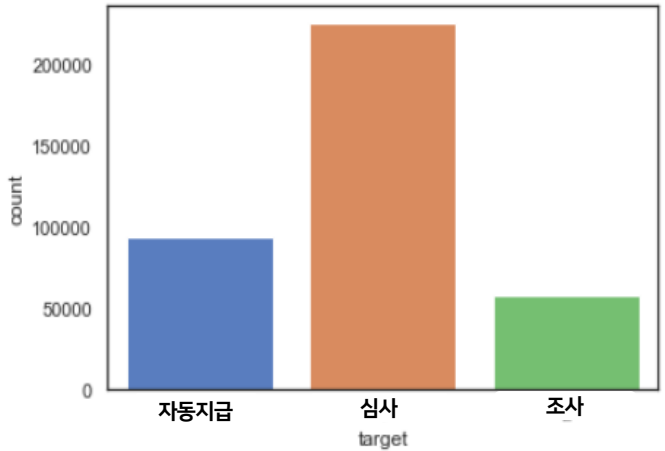
건보공단 자료에 따르면 폐렴 환자는 2014년 140만 명에서 2018년 134만 명으로 연평균 1.1%씩 꾸준히 감소했다. 국민건강보험 일산병원 호흡기내과 박선철 교수는 최근 5년간 폐렴 환자 수가 감소하고 있는 원인에 대해 "폐렴에 대한 인식의 증가와 독감예방접종이나 폐렴구균예방접종과 같은 예방 접종의 확대 등이 영향을 미쳤을 것"이라고 말했다. 월별로는 12월이 24만 명(11.8%)으로 가장 많고, 8월이 11만 명(5.2%)으로 가장 적었다. 월별 점유율 상위 5위는 12월(11.8%), 11월(10.5%), 5월(10.4%), 1월(10.2%), 4월(10%)순으로 봄(4~5월)과 겨울(12~1월), 환절기(11월)에 환자가 많았다. 계절별로는 겨울이 28.8%로 가장 높았고 여름이 18.4%로 가장 적었다. 박선철 교수는 "봄과 같은 환절기나 겨울철에는 감기나 독감과 같은 호흡기 질환이 유행하고 이로 인해 면역력이 떨어지면서 폐렴에 걸릴 위험이 높다"고 말했다.

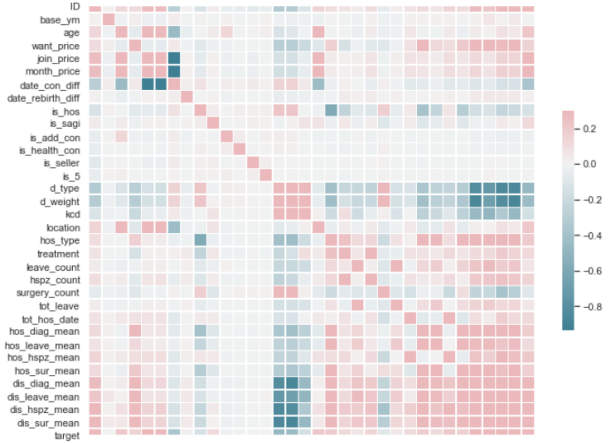
출처: 헬스조선 뉴스

2. 데이터 해석 및 가공 2-1-2. 전반적 탐색

MIRAE ASSET 미래에셋생명

Train data 377,928개 / Test data 22,072개





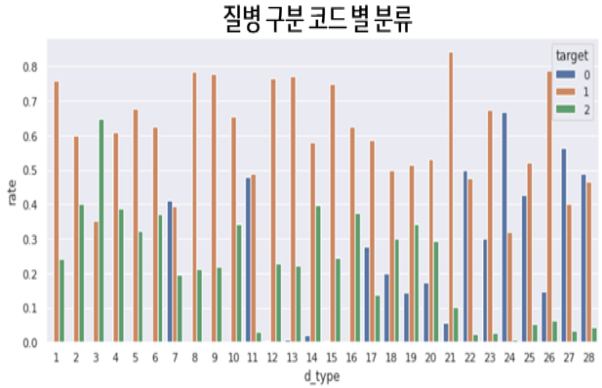
	가입금액 구간코드	보험료구간코드	지역구분코드	의료기관구분코드
max	99	99	9	9

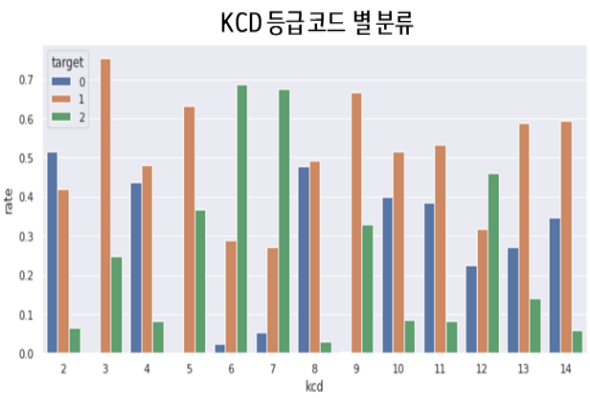
√ 범주형 변수가 많으며, Target의 분포가 비교적 **불균형**
√ 특성변수간 **상관관계** 존재
√ 결측 대신 unknown 값이 9, 99형태로 표시
√ Train data가 37만개로 충분한 상황

2. 데이터 해석 및 가공 2-2-1. 특성변수 EDA

MIRAE ASSET 미래에셋생명

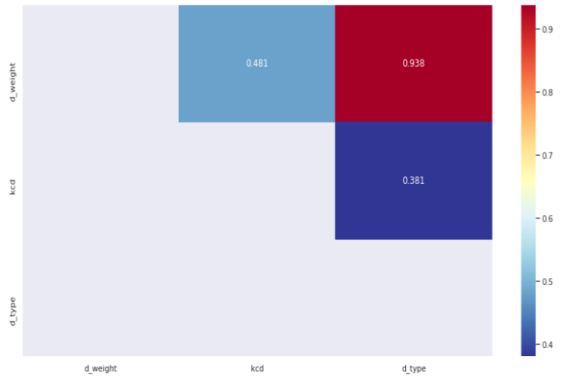
① 변수의 Target별 분포 파악 : 단계적 범주형 특성변수가 많음. 이러한 변수들에서 각 Target이 차지하는 비율을 시각화





질병구분	해당 KCD	비율
1	3	0.999427
2	4	0.853603
3	4	0.892553
5	9	0.993238
6	9	0.989145

<질병 구분 코드별 해당 KCD의 비율>



<질병 경중, KCD, 질병 구분코드 Heatmap>

질병 관련 변수

√ 타겟의 분류에 질병의 특성이 반영됨.

Ex) 질병 구분코드의 경계성(3)은 조사의 비율이 월등히 높음. 경계성은 정서 불안정, 이상 성격을 보이는 질병으로 조사의 이유가 명확.

√ 한 질병 구분코드는 동일한 KCD등급을 가질 확률이 높음.

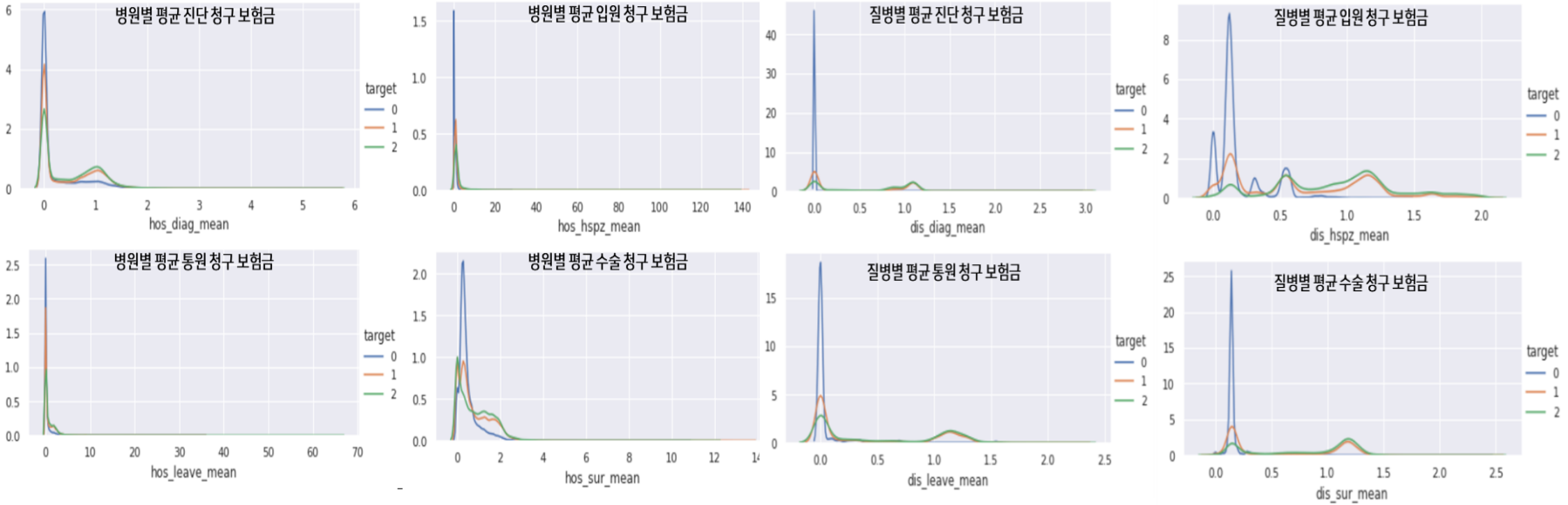
→ 이들간 상관관계가 높지만, 질병정보는 중요한 변수이므로 제거는 어려움.

2. 데이터 해석 및 가공 2-2-2. 특성변수 EDA

MIRAE ASSET 미래에셋생명

청구 보험금 관련 변수

보험사기 적발 등에 청구보험금은 매우 중요한 역할을 함.
이들을 밀도함수(KDE plot)을 통해 Target별로 시각화



√ 자동지급이 심사, 조사에 비해 두드러지는 구간이 명확함.
√ 모든 변수가 두드러지는 이상치 존재

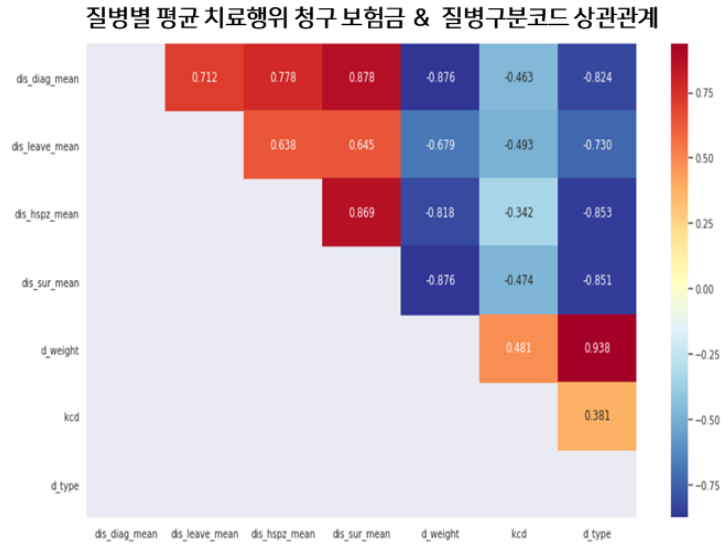
→ 이러한 구간과 이상치를 특성으로 잘 반영할 수 있는 모델 선정이 중요하다고 생각

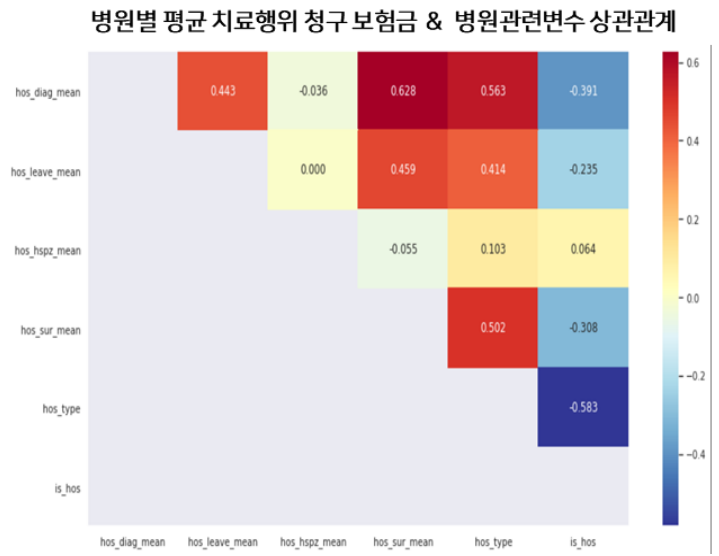
2. 데이터 해석 및 가공 2-2-3. 특성변수 EDA

MIRAE ASSET 미래에셋생명

청구 보험금 관련 변수

질병관련 변수, 병원 관련 변수들과 상관관계를 파악하기 위해 Heatmap 시각화





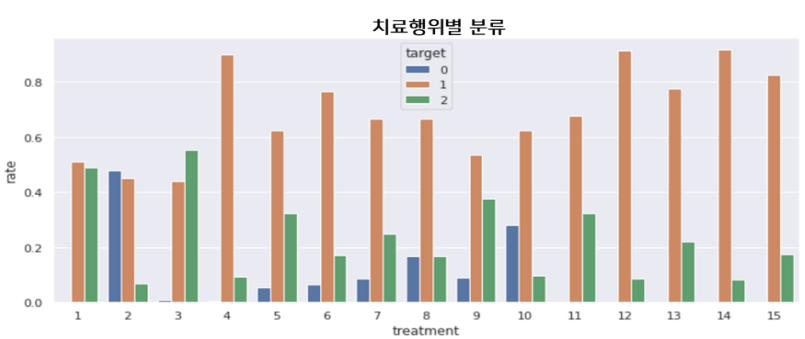
√ 질병별 청구보험금과 질병 구분코드, 병원별 청구보험금과 병원 관련변수의 상관관계 또한 두드러짐.
이들은 변수 선택으로 제거하기에 너무 중요한 변수라고 생각

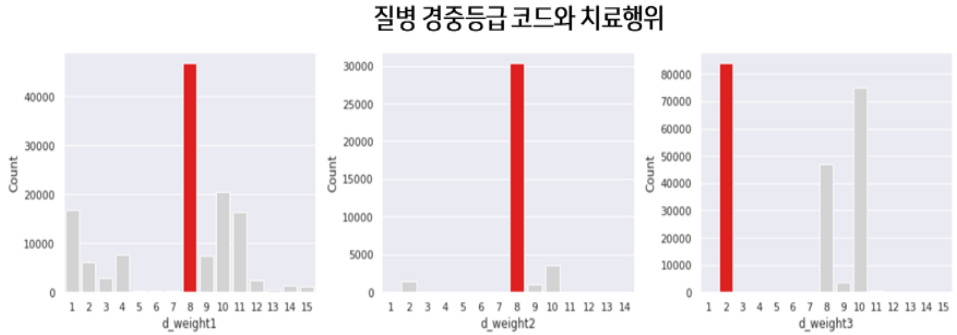
→ 다중 공선성에 영향을 받지 않는 모델 선정이 효율적일 것이라고 생각

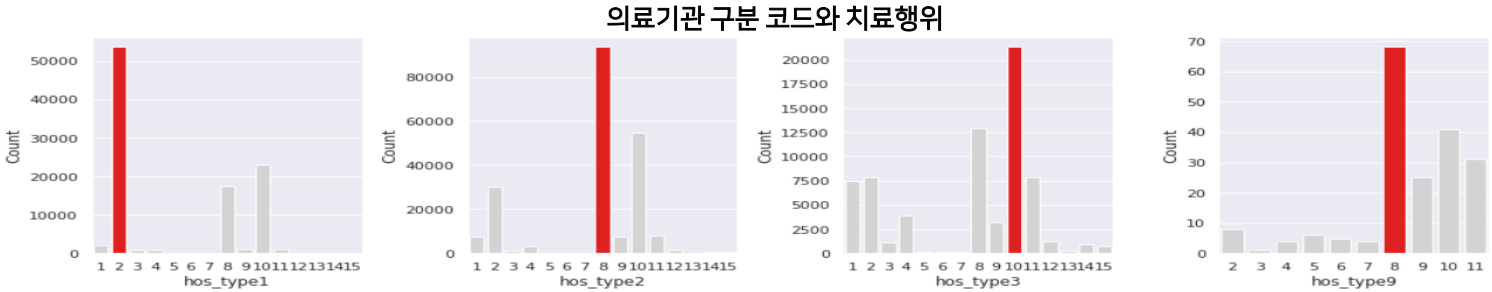
2. 데이터 해석 및 가공 2-2-4. 특성변수 EDA

MIRAE ASSET 미래에셋생명

치료행위 관련 변수







√ 질병 경중등급, 의료기관 구분의 범주별로 확연히 두드러지는 치료행위가 존재함.

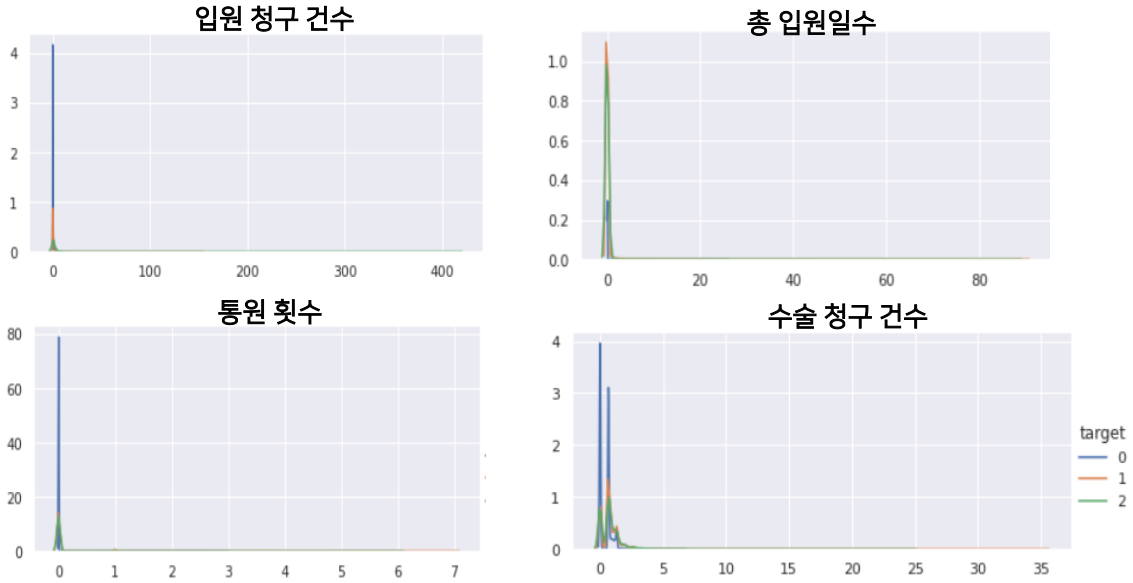
→ 치료행위를 이용한 새로운 변수 생성이 유의해 보임.

2. 데이터 해석 및 가공 2-2-5. 특성변수 EDA

MIRAE ASSET 미래에셋생명

치료행위 관련 변수

치료행위가 범주별로 두드러지는 특징을 보이므로,
치료행위와 관련된 청구 건수와 횟수도 중요한 변수일 것이라고 생각



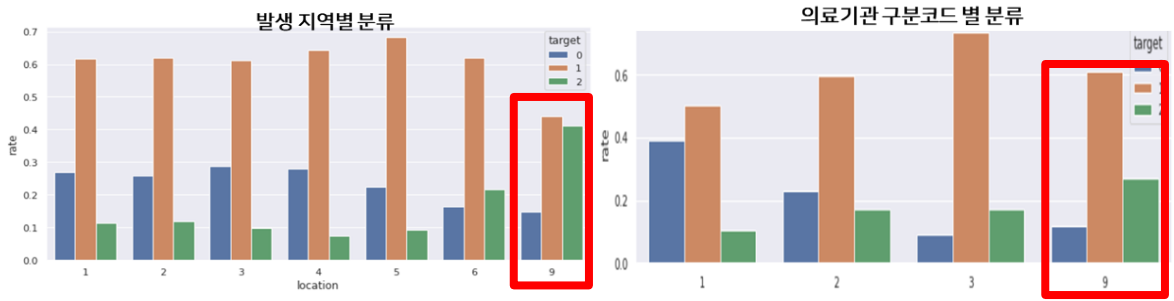
√ 청구보험금과 유사하게 자동지급이 심사, 조사에 비해 두드러지는 구간이 존재함.
√ 모든 변수가 두드러지는 이상치 존재(Righted skewed 모양)

→ 치료행위와 관련된 변수들의 이러한 특성 또한 잘 반영할 수 있는 모델이어야 함.

3) Unknown값 선별 근거

...	가입금액구간코드	보험료 구간코드	...
...	99	99	...

가입금액 구간코드와 보험료 구간코드가 모두 99인 row 34,436개 선별



Unknown의 값들은 대체로 조사의 비율이 월등히 높은 모습을 보임.

➡ 이러한 분포의 특성을 모델에 반영하고자 함

MIRAE ASSET 미래에셋생명

① 각 질병 별 청구보험금의 평균이 모든 접수년월에 상관없이 동일한지 검정

Ex) 십이지장궤양(14) 데이터

…	질병 구분 코드	접수년월	질병별 통원 청구보험금	…
…	14	201901	1.2002	…
	14	201903	1.1708	
	…	…	…	

D(i) : i월의 질병별 통원 청구 보험금의 평균

H_0 : $D(1) = D(2) = \cdots = D(11)$

H_1 : At least one not holds

	DF	Sum sq	Mean sq	F value	P-value
접수년월	10	35.47	3.547	25.45	<2e-16
Residuals	196	27.32	0.139		

P-value가 매우 작으므로 유의수준 0.01하에서 귀무가설을 기각함
즉, 십이지장궤양에서 질병별 통원 청구 보험금의 모평균은 월별로 모두 같지 않음
다른 질병에서도 이 귀무가설은 모두 기각됨

② 특정 월의 각 질병별 청구보험금의 평균이 직전 월과 동일한지 검정

① 에서 귀무가설을 기각함으로써 질병별 청구보험금의 모평균이 모두 같지 않음을 확인

But, 어느 월의 차이인지는 정확히 알 수 없음
따라서 구체적인 월별 차이를 파악하기 위한 사후 검정이 필요함

특정 월의 바로 직전 월은 질병 트렌드에 비슷한 영향을 받아
질병별 청구보험금의 평균이 유사할 것이라는 가설을 설정

특정 월의 각 질병별 청구보험금의 평균이
그 직전 월과 동일한지 일대일 검정

(모든 월을 특정 월로 놓고 pairwise 검정 진행)

② 특정 월의 각 질병별 청구보험금의 평균이 직전 월과 동일한지 검정

Ex1) 같은 십이지장궤양(14) 데이터의 특정 월을 11월로 선정한 후, 10월과 11월의 동일성 검정

	DF	Sum sq	Mean sq	F value	P-value
접수년월	1	0.999	0.999	4.167	0.0529
Residuals	23	5.514	0.2398		

H0 : D(10) = D(11)
H1 : At least one not holds

Ex 2) 같은 십이지장궤양(14) 데이터의 특정 월을 10월로 선정한 후, 9월과 10월의 동일성 검정

	DF	Sum sq	Mean sq	F value	P-value
접수년월	1	0.504	0.5038	1.744	0.194
Residuals	40	11.555	0.2889		

H0 : D(9) = D(10)
H1 : At least one not holds

P-value가 0.01보다 크므로 유의수준 0.01하에서 귀무가설을 인용함
즉, 십이지장궤양에서 질병별 통원 청구 보험금의 모평균은 직전월과 유사함
다른 질병과 월에서도 이러한 특성이 대부분 나타남

단, 같은 십이지장궤양(14) 데이터에 대해 9, 10, 11월 3개의 월을 검정하면
0.01유의수준에서 귀무 가설을 기각함

	DF	Sum sq	Mean sq	F value	P-value
접수년월	2	4.116	2.0582	7.249	0.00016
Residuals	53	15.049	0.2839		

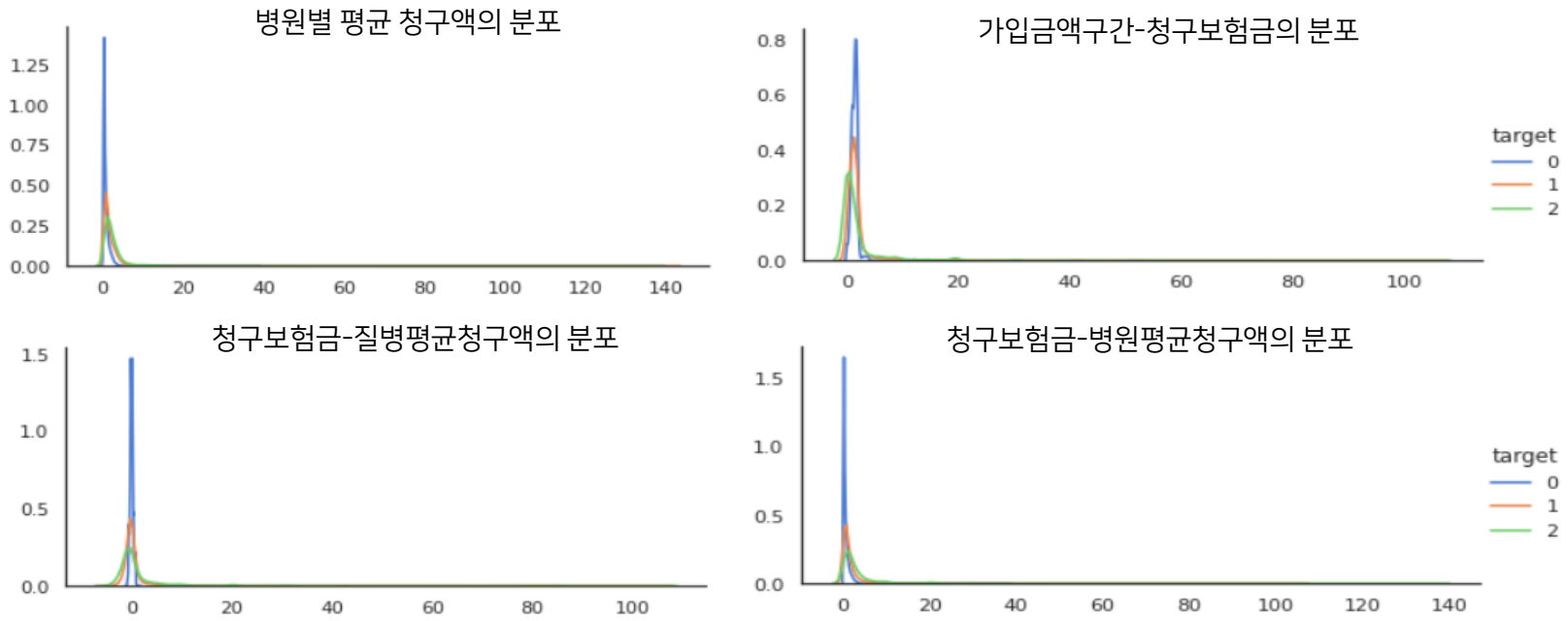
H0 : D(9) = D(10) = D(11)
H1 : At least one not holds

즉, 9월 10월 11월의 세 집단을 비교했을 때는 모평균이 같지 않으나,
각각 직전 월만 비교했을 때는 모평균이 유사함을 추론할 수 있음

2. 데이터 해석 및 가공 2-2-11. 특성변수 EDA

MIRAE ASSET 미래에셋생명

[청구보험금과 관련한 파생변수들의 분포]



0에 가까울수록 <u>평균과 비슷하다는 의미로 해석</u> 가능

→ 모든 분포가 0에 가까울수록 자동지급의 밀도가 높음
이는 자동지급일수록 청구보험금은 평균과 유사함(비이상적임)을 뜻함

6 치료행위코드

코드	치료행위				설명
	입원	통원	수술	진단	
1				Y	질병 진단만 받음
2			Y		수술치료만 진행
3			Y	Y	질병 진단 받고 수술치료 진행
4		Y			통원치료만 진행
5		Y		Y	질병 진단 받고 통원치료 진행
6		Y	Y		수술치료 후 통원치료 진행
7		Y	Y	Y	진단 받고, 수술도 받고, 통원치료도 받음
8	Y				입원치료만 진행
9	Y			Y	질병 진단 받고 입원치료 진행
10	Y		Y		입원 및 수술치료 진행
11	Y		Y	Y	질병 진단 받고 입원 및 수술치료 진행
12	Y	Y			입원 및 통원치료 진행
13	Y	Y		Y	질병 진단 받고 입원 및 통원치료 진행
14	Y	Y	Y		입원, 수술, 통원치료 모두 진행
15	Y	Y	Y	Y	질병 진단 받고 입원, 수술, 통원치료 모두 진행

치료행위코드

√ 과제설명자료의 [별첨3] 을 활용하여 입원, 통원, 수술, 진단이라는 4가지 범주를 만듦

√ 해당 치료행위코드에 대응하는 치료행위 범주에 1을 부여

Ex) 치료행위코드 5 :

입원	통원	수술	진단
0	1	0	1

MIRAE ASSET 미래에셋생명

청구보험금 : EDA에서 청구보험금에 대한 중요성을 인지하였고,
: 이를 다양한 관점에서 반영하고자 함

Feature 01 **병원별 평균 청구액**

치료행위코드에 대응하는 병원별 평균 입원, 통원, 수술, 진단액의 합을 반영

Feature 02 **질병별 평균 청구액**

치료행위코드에 대응하는 질병별 평균 입원, 통원, 수술, 진단액의 합을 반영

MIRAE ASSET 미래에셋생명

Feature 03 가입금액 구간별 평균 청구금액 - 청구 보험금

Groupby를 통해 train의 가입금액 구간별 평균 청구 보험금을 계산 후
| 가입금액별 평균 청구 보험금 – 청구 보험금 | 수식 생성, 적용

Feature 04 청구 – 질병 평균 청구액

청구보험금과 해당 질병의 평균적인 청구보험금의 차이를 반영

Feature 05 청구 - 병원 평균 청구액

청구보험금과 병원의 평균적인 청구보험금의 차이를 반영

∴ 평균 청구 보험금과의 차이를 통해 비이상적 상황(보험사기 등)을 반영하고자 함

MIRAE ASSET 미래에셋생명

기타 : 고객 관련 변수 생성

Feature 01 재가입여부

보험료를 납부하지 않고 일정 기일이 경과하면 그 계약은 해지되는데,
이를 다시 회복한다면 보험금을 받고자 부활한 것일 수도 있다고 가정

Feature 02 총 청구건수

데이터에 있는 청구 건수를 더해 각 고객이 얼마나 많은 청구가 있었는지 반영

Feature 03 통원 + 입원일수

얼마나 많은 통원과 입원행위가 있었는지를 반영함

MIRAE ASSET 미래에셋생명

모델 선택 근거

① 데이터의 구간별 특징을 잘 반영할 수 있는 모델이어야 함

- EDA결과, target별로 데이터의 구간별 특징이 두드러짐

  예시 1) 질병 경중등급이 높은 Data는 자동지급이 관측되지 않음

  예시 2) 청구보험금이 0인 경우 대부분 조사가 이루어짐
  반대로 청구보험금이 다른 값에 비해 높다면
  (train data에서는 2.1505보다 크다면) 자동지급이 관측되지 않음

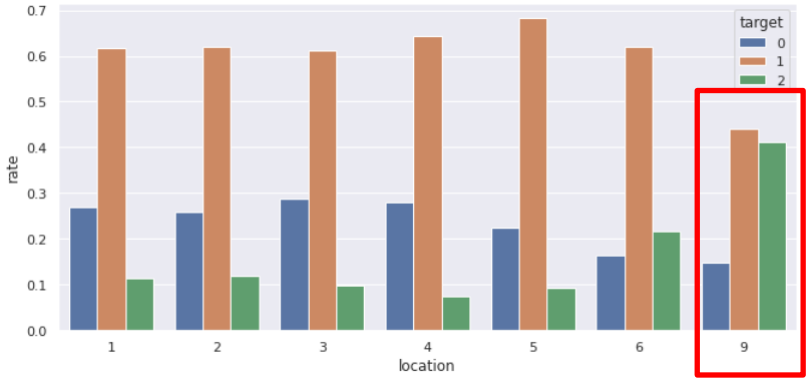
- Unknown 데이터를 모델에 반영하기 위해 구간을 나눠
  해당 데이터를 범위에서 벗어난 하나의 특성으로 반영해야 함 ex) 1, 2, 3, 9

MIRAE ASSET 미래에셋생명

모델 선택 근거

② Unknown 데이터를 제거하지 않고 모델에 반영

- Unknown 데이터가 다른 데이터에 비해 target이 눈에 띄는 특징들을 보였고, 이를 모델에 잘 반영할 수 있어야 함.



<발생 지역별 분류>

	month_price	target	보험료구간코드별 target의 수
30	11	0	810
31	11	1	1746
32	11	2	29
33	99	0	1519
34	99	1	10367
35	99	2	22550

<보험료구간코드별 분류>

MIRAE ASSET 미래에셋생명

모델 선택 근거

③ 정확도, 학습 속도, 예측시간이 동시에 고려되어야 함

→ 대표적인 머신러닝 알고리즘으로 세 요소를 종합적으로 평가(튜닝은 하지 않음)

[No tuning]	학습시간(초)	예측시간(초)	F1_score
Logistic	8.54	0.008	0.418
Decision Tree	1.62	0.01	0.626
SVM	20270.68	0.0079	0.29

-학습 속도, **F1_score**에서 **Decision Tree**가 우수한 성능을 보였음

-**Logistic**은 특히 자동지급을 잘 분류하지 못했고. **SVM**은 특히 심사를 잘 분류하지 못하였음

MIRAE ASSET 미래에셋생명

모델 계열 선택 : Decision Tree

→ Tree의 강점을 살린 Ensemble Learning을 활용

- Ensemble : 하나의 데이터를 여러 학습 모델에 학습시키고, 학습결과를 결합
  → 과적합 방지, 정확도 향상효과
- Tree계열 앙상블 모델(Random Forest, ExtraTree, LightGbm) 비교 검증

		RF	Extra Tree	LightGbm
공통점		트리 기반의 앙상블 모델 → 많은 트리를 생성해 그들의 예측값을 보팅(voting)		
차이점	샘플링 기법	부트스트랩 (Bootstrap, 복원추출)	샘플링 X	GOSS (정보획득 큰 데이터 샘플링)
	분할 지점 설정	최적의 분할지점	랜덤하게 선택	Pre-sorted (사전 정렬하고 모든 변수 계산)

검증 목표

√ 직전 달만 사용하여 target을 예측하는 것이 성능향상에 도움이 되는지 검증

√ 최적의 Ensemble algorithm은 무엇인지 검증

√ 통계적 검정으로 밝힌 기간에 대한 가설이 모델 적용에도 성립하는지에 대한 검증을 목표로 평가 기준(F1 Score와 시간)을 설정하여 우리의 분석 방향에 맞춰 경험적 검증을 진행

3. 모델링 3-2-2. 검증(F1 score)

MIRAE ASSET 미래에셋생명

F1_score	1개월	3개월	6개월	10개월
Extra Tree	0.83	0.816	0.81	0.803
Random Forest	0.806	0.793	0.791	0.787
Light GBM	0.797	0.776	0.765	0.762

① F1 score

1. Extra Tree
2. Random Forest
3. Light GBM

3. 모델링 3-2-3. 검증(시간)

MIRAE ASSET 미래에셋생명

학습 시간(초)	1개월	3개월	6개월	10개월
Extra Tree	5.6	22.9	58.5	119.5
Random Forest	10.4	39.3	91.5	173.67
Light GBM	7.337	13.148	22.17	33.484

예측 시간(초)	1개월	3개월	6개월	10개월
Extra Tree	0.9	1.13	1.47	1.69
Random Forest	0.78	1	1.14	1.35
Light GBM	0.88	0.87	0.82	0.88

② 학습 시간

(1 개월 기준)

1. Extra Tree
2. Light GBM
3. Random Forest

③ 예측 시간

(1 개월 기준)

1. Random Forest
2. Light GBM
3. Extra Tree

MIRAE ASSET 미래에셋생명

모델 선택 : Extra Tree (Extremely randomize tree)

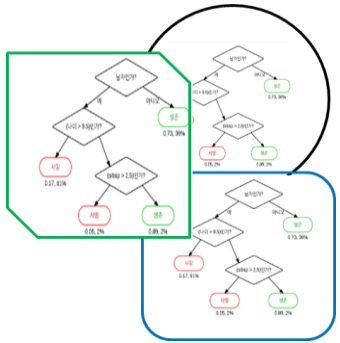
1) 알고리즘

√ Random Forest를 기반으로
더욱 랜덤성을 강화한 트리기반 앙상블 모델

➡ 노드 분할 지점(Cut point)과 사용할 feature를 랜덤하게 선택
(Explicit randomization)

√ 노드를 완전히 무작위로 분할
(choosing cut-points fully at random)

√ 부트스트랩(복원추출)을 하지 않고 모든 샘플을 사용



모델 선택 : Extra Tree

2) 장점

√ Explicit randomization 기법으로 분산 감소 효과
+ 노이즈가 있는 Feature에 잘 대응

√ 부트스트랩을 사용하지 않고 전체 데이터셋을 모두 반영함으로써 편향을 줄이고 일반화된 성능을 얻을 수 있음

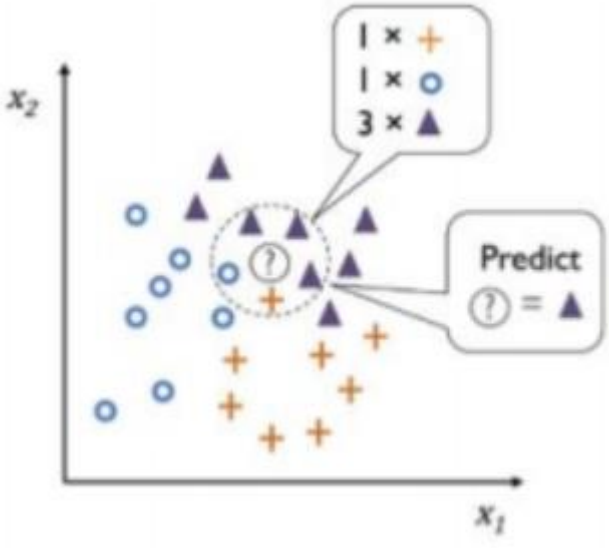
√ 노드 분할지점(Cut-point)를 최적화하지 않기때문에 계산 복잡도를 줄이고 학습 시간을 대폭 단축

단, 랜덤성이 강화된 알고리즘이기에 Random state에 지나치게 의존될 가능성이 있음

→ 모델 발전 및 고도화 과정에서 이를 보완하고자 함

모델 고도화 ① Extra tree의 랜덤성 보완

사례 기반 모델 : KNN Classifier (K-최근접 이웃 알고리즘)



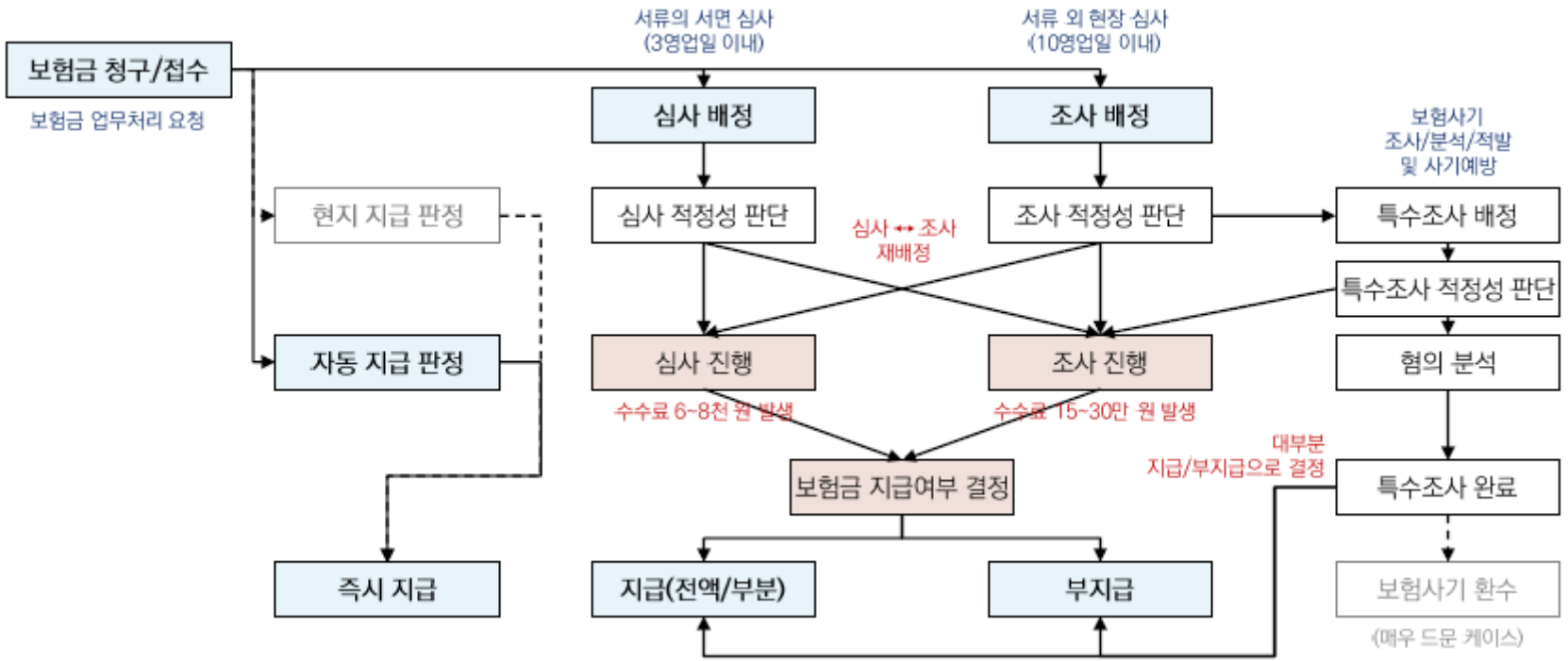
√ 기존 데이터와 가까운 이웃의 정보로
새로운 데이터를 예측하는 방법론

√ Train data 자체가 모형이 되는 효과
= 사례 기반 모델(Instance based learning)

규칙기반 모델인 Extra Tree에 결합함으로써
전반적인 모델의 안정성을 높임

MIRAE ASSET 미래에셋생명

심사/조사 Process



MIRAE ASSET 미래에셋생명

새로운 평가지표 제안

목적 : 자동지급에 대한 Precision을 집중하면서 전반적인 성능도 고려
극단적으로 모델이 조사와 심사로만 분류하는 상황 방지

(0.7 * 자동지급의 precision score) + (0.3 * 전체 class의 F1 score 평균)

F1 score : 0.84308 | Precision : 0.77782
new_score : 0.797398

2 Stage model

MIRAE ASSET 미래에셋생명

Collaborative Filtering

많은 사용자들로부터 얻은 기호정보에 따라 **사용자들의 관심사들을 자동적으로 예측**하게 해주는 방법

특정 사용자의 정보에만 국한된 것이 아니라 많은 사용자들로부터 수집된 정보를 사용

고객의 선호도와 관심 표현을 바탕으로 선호도, 관심에서 **비슷한 패턴을 가진 고객들을 식별**해 내는 기법

비슷한 취향을 가진 고객들에게 서로 아직 구매하지 않은 상품들은 교차 추천이나
분류된 **고객의 취향이나 생활 형태에 따라 관련 상품을 추천**하는 형태의 서비스를 제공하기 위해 사용된다.

-위키백과-

User Based Collaborative Filtering

$$r_{xi} = \frac{\sum_{y \in N} s_{xy} r_{yi}}{\sum_{y \in N} s_{xy}}$$

r_{xi} : 사용자 x가 아이템 i 구매 유무
N : 아이템 i를 평가한 사람 중 나와 가장 유사한 k명의 사용자 집합
s_{xy} : 사용자 x와 사용자 y의 유사도

User 정보를 활용하여 **User간 유사도를 측정**한 후 유사도가 높은 User에 높은 Weight를 낮은 User에 낮은 Weight를 삽입하여
기존에 구매하지 않은 새로운 상품을 추천하는 방식이다.

Item Based Collaborative Filtering

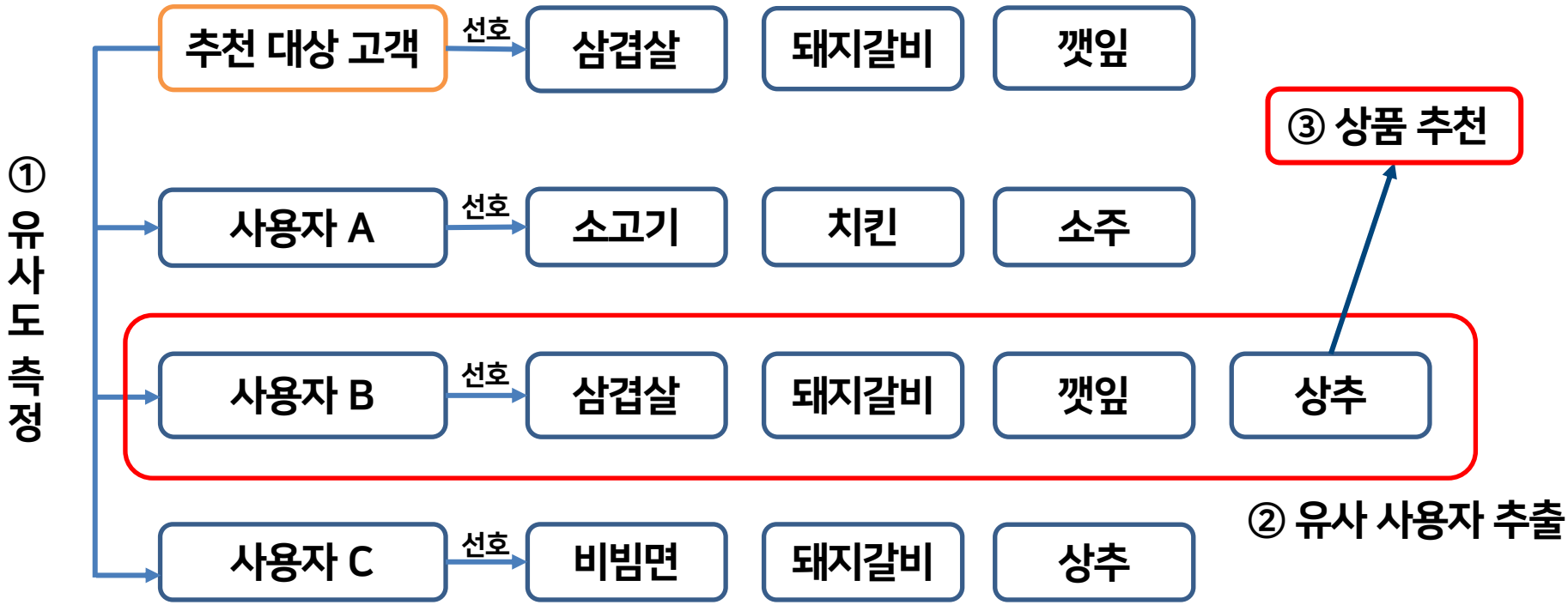
$$r_{xi} = \frac{\sum_{j \in N} s_{ij} \times r_{xj}}{\sum_{j \in N} s_{ij}}$$

r_{xi} : 사용자 x가 아이템 i 구매 유무
N : 평가한 아이템 중 아이템 i와 가장 유사한 k개의 아이템 집합
s_{ij} : 아이템 i와 아이템 j의 유사도

Item 정보를 활용하여 **Item간 유사도를 측정**한 후 유사도가 높은 Item에 높은 Weight를 낮은 Item에 낮은 Weight를 삽입하여
기존에 구매하지 않은 새로운 상품을 추천하는 방식이다.

MIRAE ASSET 미래에셋생명

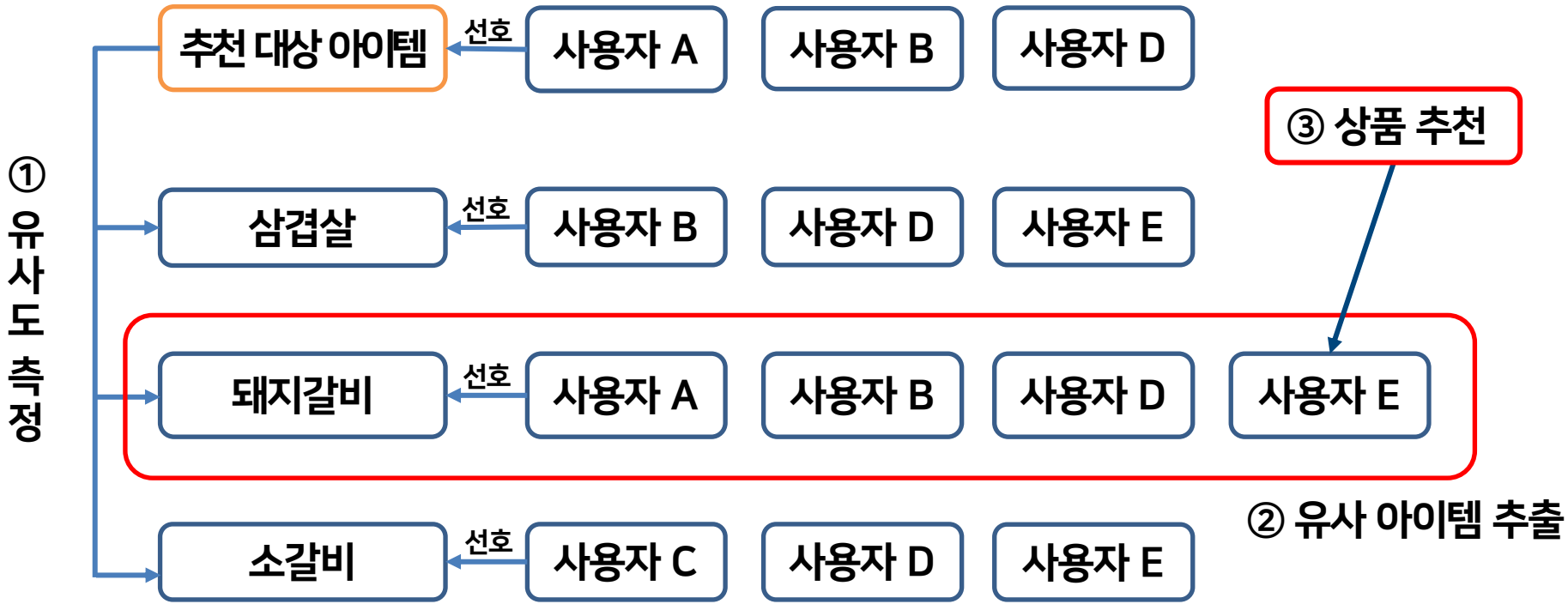
User Based Collaborative Filtering



추천 대상 고객과 가장 유사한 사용자 B를 선택한후 사용자 B의 Item 중
추천 대상 고객의 Item에서 없는 상품인 상추를 추천한다.

MIRAE ASSET 미래에셋생명

Item Based Collaborative Filtering



추천 대상 아이템을 기준으로 선호 유저가 유사한 아이템을 선정한 뒤

해당 아이템을 구매 하였으나 추천 대상 아이템을 구매하지 않은 사용자 E에게 추천