LOVE AGE

MIRAE ASSET 미래에셋생명

미래에셋 생명
금융 빅데이터 페스티벌

팀 막고라

LOVE AGE

미래에셋생명 빅데이터 페스티벌

CONTENTS

0. 팀 소개

MIRAE ASSET 미래에셋생명

팀장

문성민
빅데이터경영통계학과 4학년
010-2355-2369
tjdals0410@kookmin.ac.kr

김태현
빅데이터경영통계학과 4학년
010-2929-5390
hyoun3024@kookmin.ac.kr

이다은
통계학과 2학년
010-5093-1892
goodgpt@korea.ac.kr

팀 막고라

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 보험사기 예방

√ 데이터를 통해
보험사기를 미리 예측 가능

출처 연합뉴스

② 자동 심사 가능

√ 기존 심사 인력의
효율적 배치 가능
→ 기존 인력을 까다로운 심사에
투입하여 더 세부적인 심사 가능

√ 기계학습 모델을 활용한
고객 만족도를 높일 수 있는
다양한 서비스 구축 가능

MIRAE ASSET 미래에셋생명

보험 청구 적정성 판단에 왜 기계학습이 필요할까?

기계학습을 인간이 탐색하지 못하는 패턴을 발견하는데 용이

보험사기로 인한 손실 예방

+

효율적인 인력 운용 가능

+

고객에게 질 높은 보험 서비스 제공

① 보험사기 예방

√ 데이터를 통해
보험사기를 미리 예측 가능

② 자동 심사 가능

√ 자동 심사 인력의
효율적 배치 가능
→ 기존 인력을 까다로운 심사에
투입하여 더 세부적인 심사 가능

√ 기계학습 모델을 활용한
고객 만족도를 높일 수 있는
다양한 서비스 구축 가능

출처 연합뉴스

MIRAE ASSET 미래에셋생명

두 이점을 극대화할 수 있는 분석 방향 설정

① 보험사기 예방

√ 청구 보험금을 중심으로
이상 징후를 잘 잡아낼 수 있는
특성변수 생성
(In Feature Engineering)

② 자동 심사 가능

√ 경험기반모델 + 사례기반모델

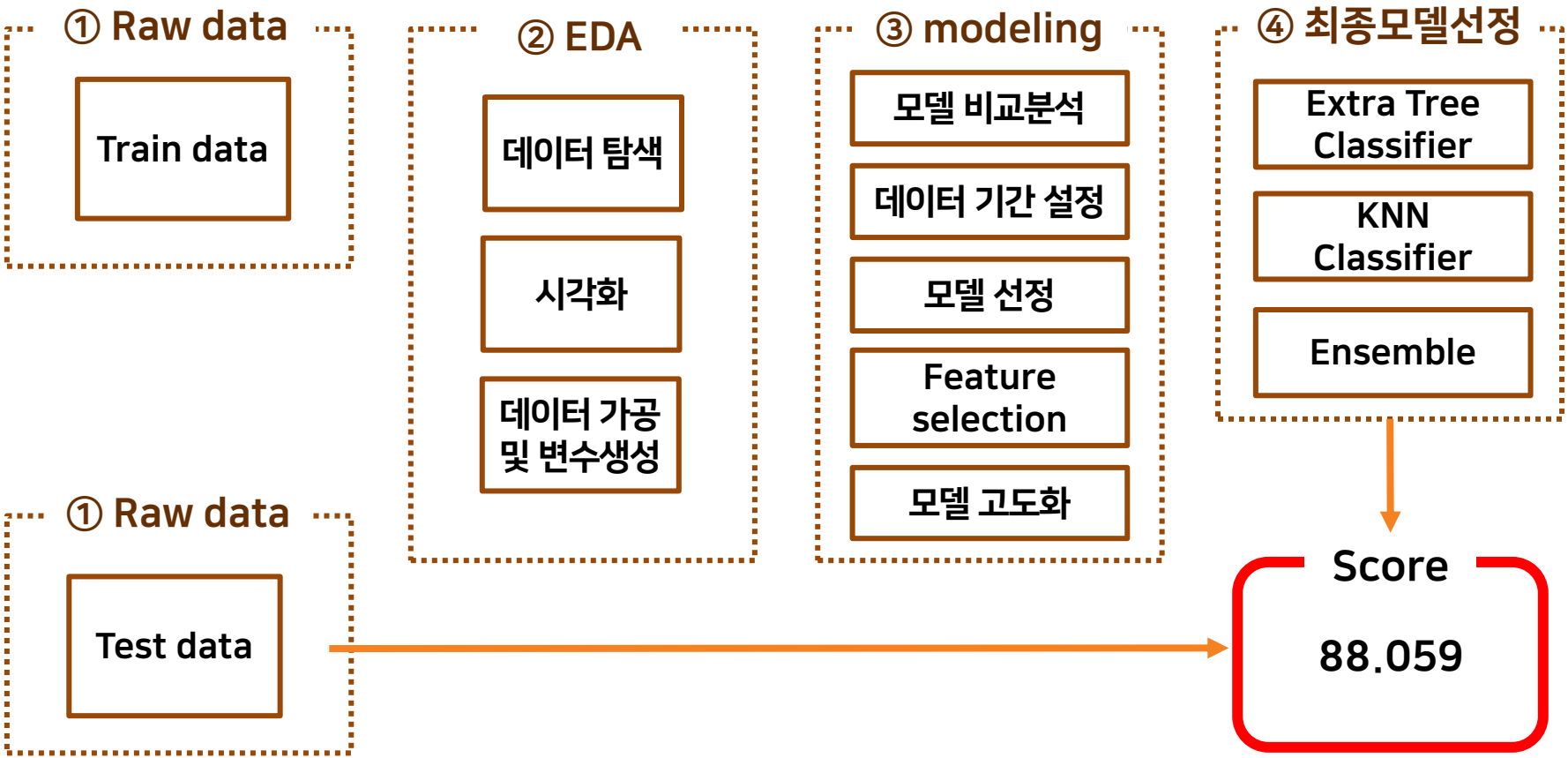
√ 현업 활용 가능성을 고려한
정확한 모델 지향
(In Modeling)

각 분야별 시스템 활용 방안 제안

1. 개요

MIRAE ASSET 미래에셋생명

전체 분석 프로세스



LOVE AGE

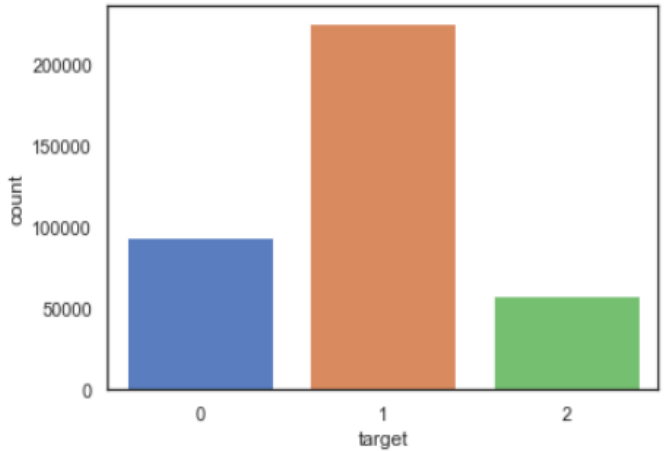
2. 데이터 해석

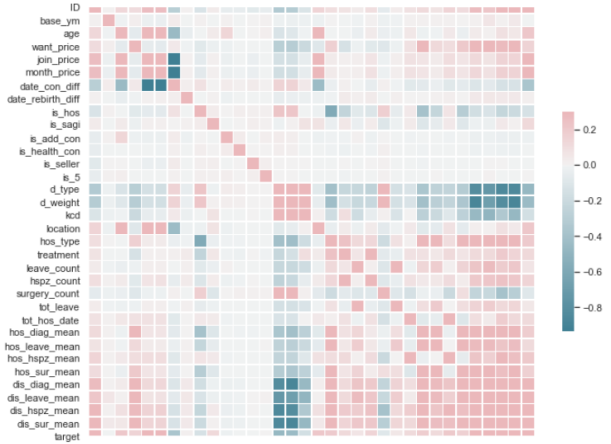
2-1. 전반적 탐색
2-2. 특성변수 EDA

2. 데이터 해석 2-1. 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 2-1. 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

√ **unknown**만의 분포 파악 후, 처리 방안 고안

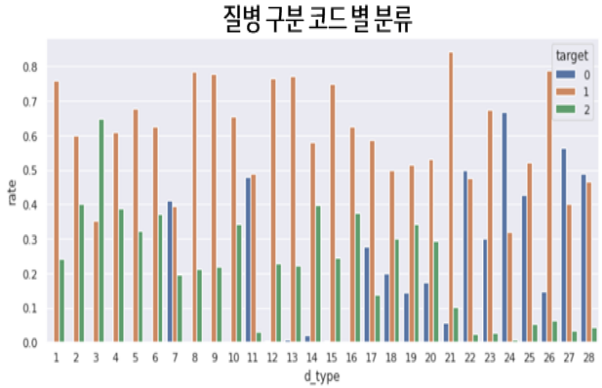
√ Train data의 경향성 파악 후 **데이터 정제** 방안 고안

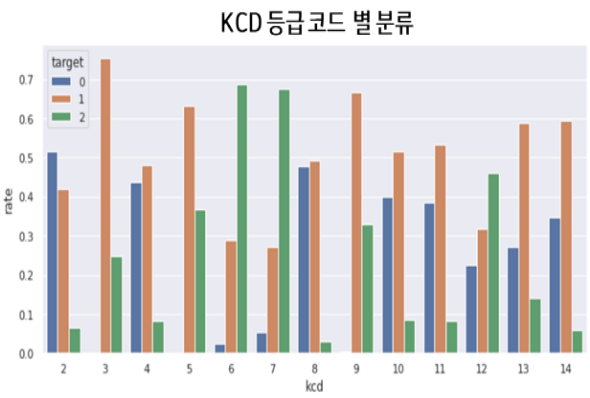
이들은 **모델 선정의 근거**로 작용

2. 데이터 해석 2-2. 특성변수 EDA

MIRAE ASSET 미래에셋생명

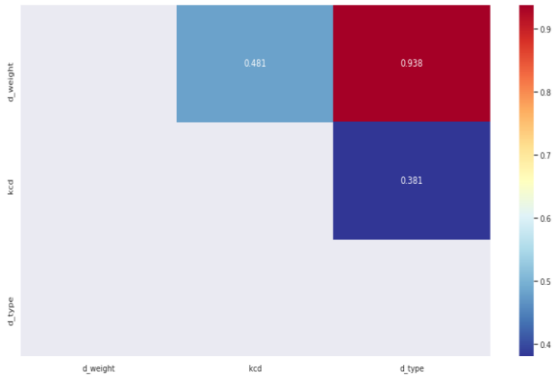
① 변수의 Target별 분포 파악 : 단계적 범주형 특성변수가 많음. 이러한 변수들에서 각 Target이 차지하는 비율을 시각화





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.

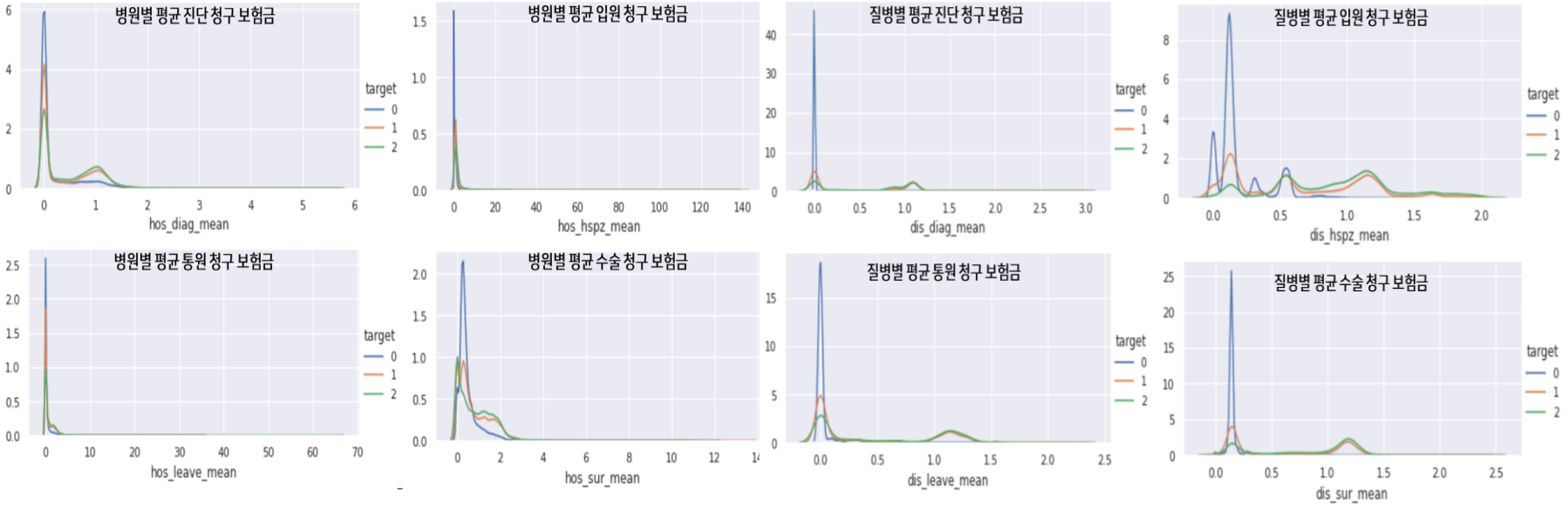
→ 이들간 상관관계가 높지만, 질병정보는 중요한 변수이므로 제거는 어려움.

2. 데이터 해석 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



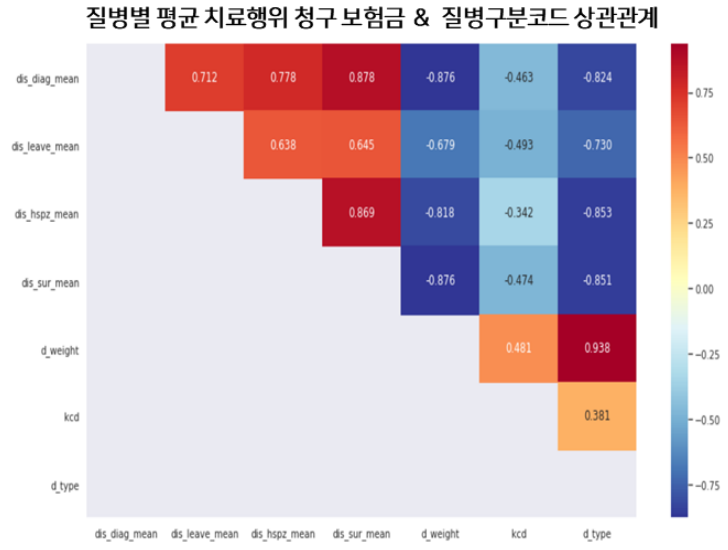
√ 자동지급이 심사, 조사에 비해 두드러지는 구간이 명확함.
√ 모든 변수가 두드러지는 이상치 존재

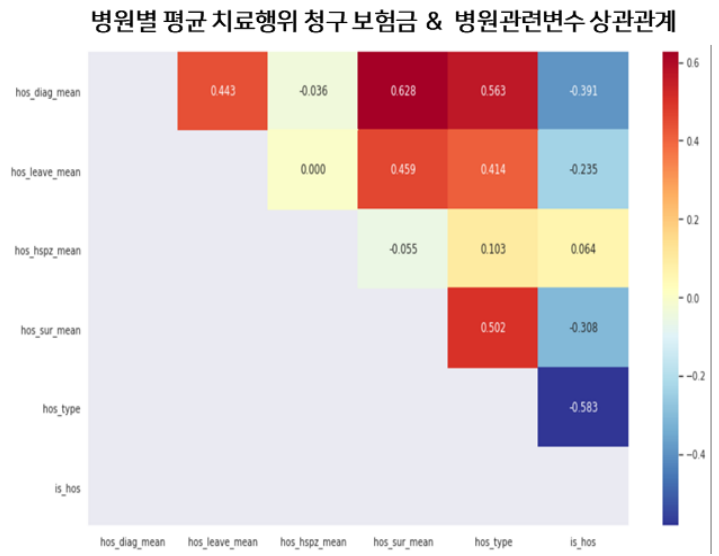
→ 이러한 구간과 이상치를 특성으로 잘 반영할 수 있는 모델 선정이 중요하다고 생각

MIRAE ASSET 미래에셋생명

청구 보험금 관련 변수

질병관련 변수, 병원 관련 변수들과 상관관계를 파악하기 위해 Heatmap 시각화





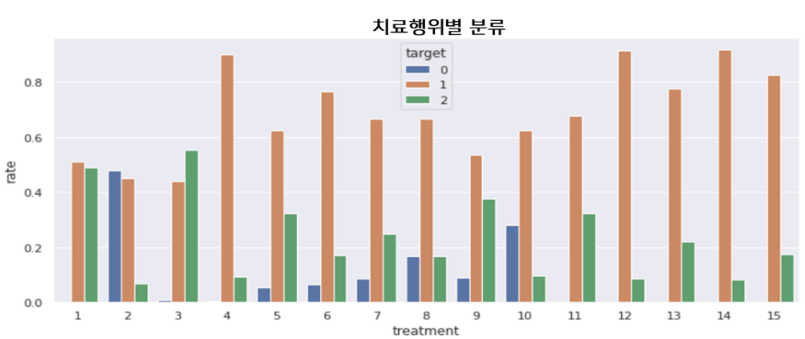
√ 질병별 청구보험금과 질병 구분코드, 병원별 청구보험금과 병원 관련변수의 상관관계 또한 두드러짐.
이들은 변수 선택으로 제거하기에 너무 중요한 변수라고 생각

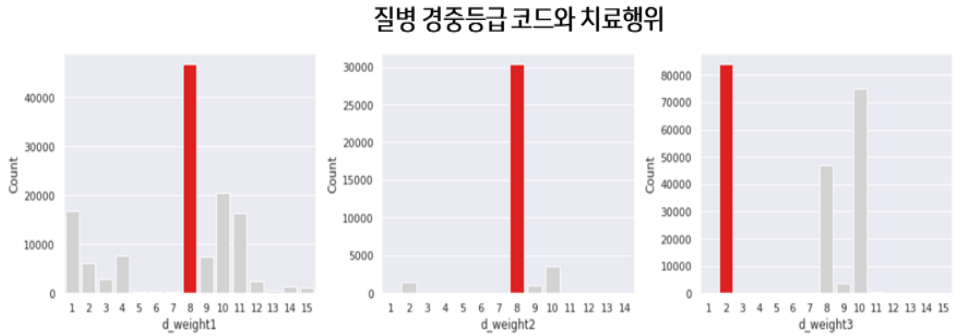
→ 다중 공선성에 영향을 받지 않는 모델 선정이 효율적일 것이라고 생각

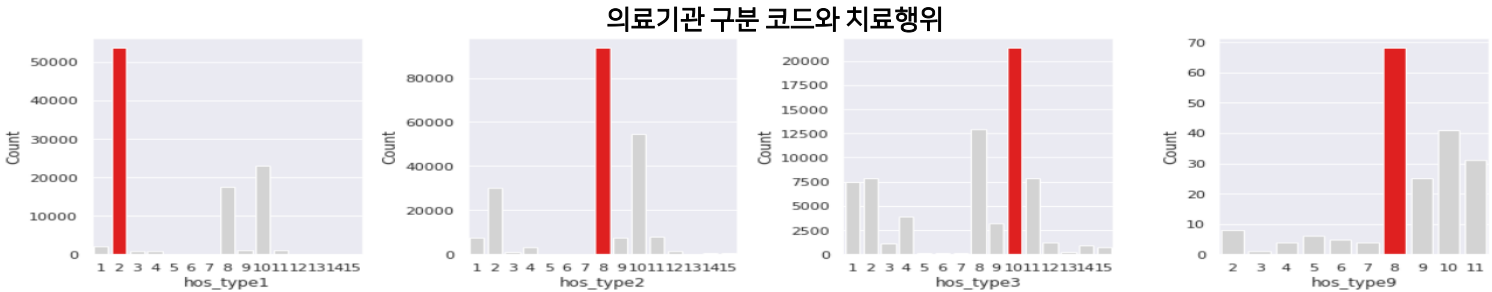
2. 데이터 해석 2-2. 특성변수 EDA

MIRAE ASSET 미래에셋생명

치료행위 관련 변수







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

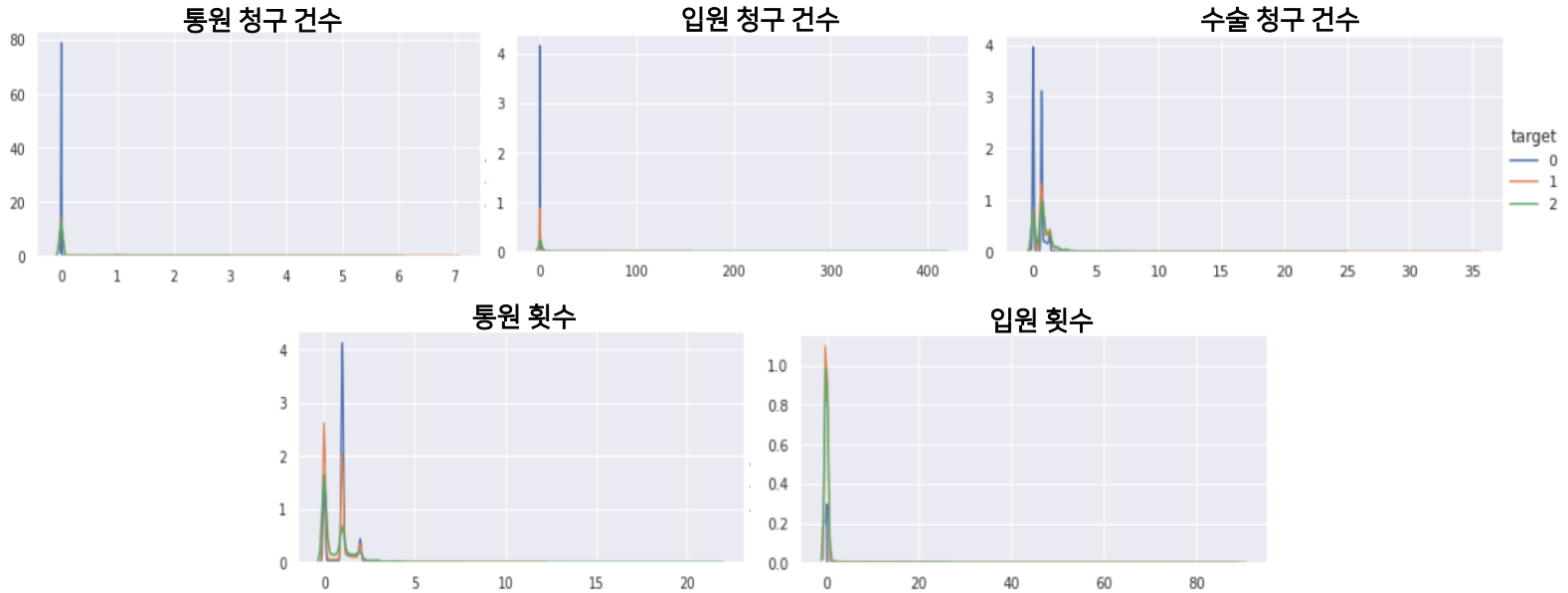
→ 치료행위를 이용한 새로운 변수 생성이 유의해 보임.

2. 데이터 해석 2-2. 특성변수 EDA

MIRAE ASSET 미래에셋생명

치료행위 관련 변수

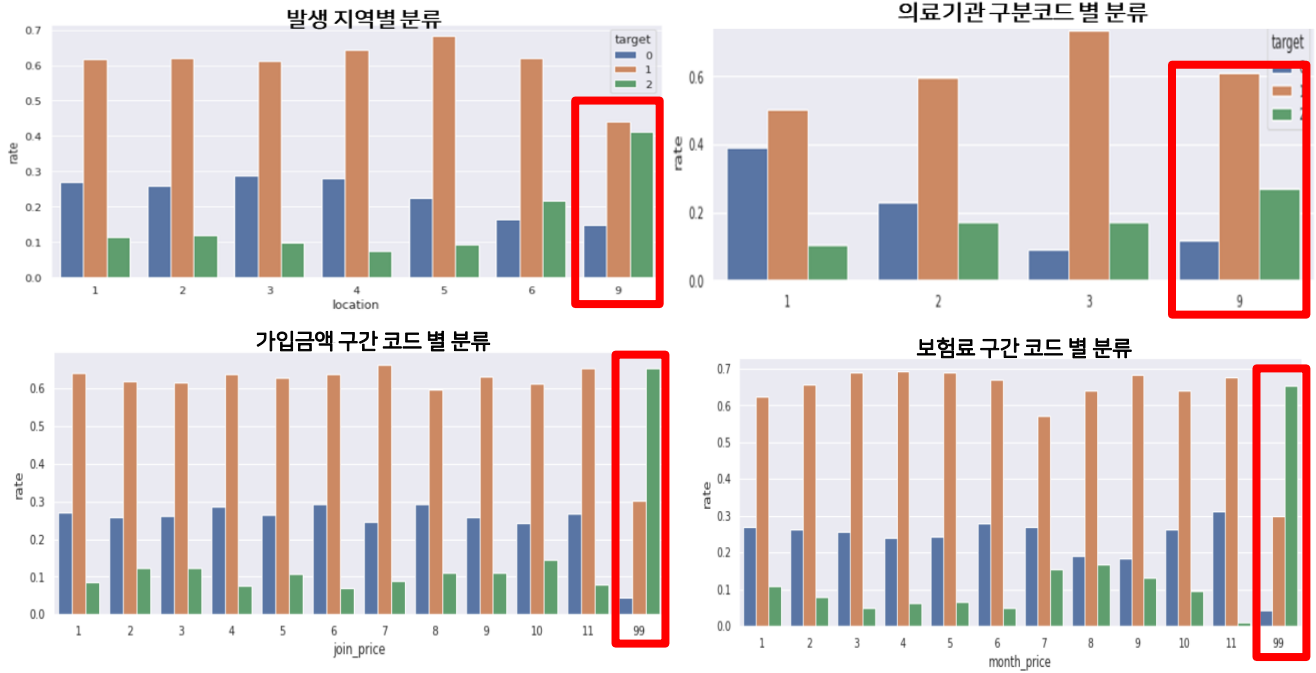
치료행위가 범주별로 두드러지는 특징을 보이므로,
치료행위와 관련된 청구 건수와 횟수도 중요한 변수일 것이라고 생각



√ 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.

② 변수별 Unknown값의 분포 파악 : Unknown만의 분포를 파악하고, 이에 대응하는 것이 중요할 것이라고 생각

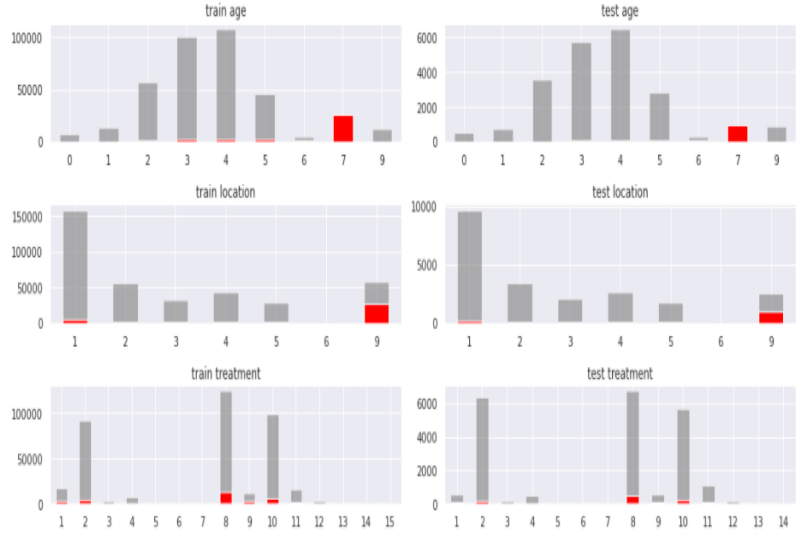


√ Unknown의 값들은 조사의 비율이 월등히 높음.
√ 가입금액구간 Unknown에 해당하면, 보험료 구간도 Unknown에 해당.
이 외에도 이들은 여러 공통된 특징을 가짐.

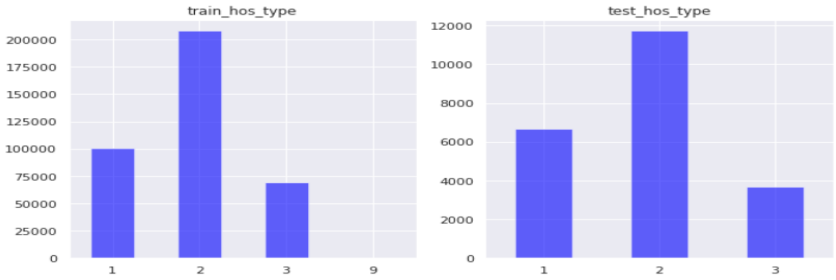
2. 데이터 해석 2-2. 특성변수 EDA

MIRAE ASSET 미래에셋생명

Train, test의 고객 나이, 지역, 치료 행위에서 차지하는 Unknown의 비율 비교



의료기관 구분코드의 Train/Test 분포 비교 (Test에는 9가 존재하지 않음)

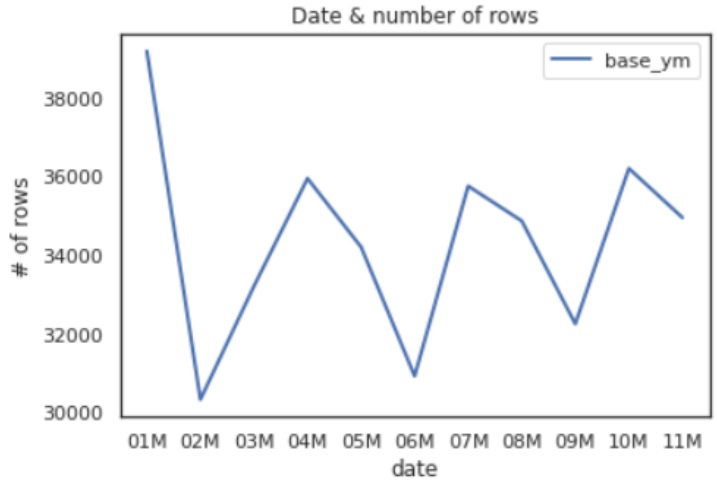


√ Q&A 게시판 질문 결과,
이들은 보험금이 지급되지 않거나,
비식별화 조치 중 생성된 값일 수 있다는
정보를 얻음.

√ Unknown값은 Train과 Test에
비슷하게 분포하고 있음.

√ 예외적으로, 의료기관 구분코드의
Unknown 범주(9)는
Test에 존재하지 않음.
하지만 이러한 사실을 모델에 반영하는 것은
현업의 관점에서 Data leakage이므로,
Train의 unknown을 따로 제거하지 않음.

③ Train의 시간적 경향성 파악: Test set이 12월의 데이터이기 때문에, 월별 특성이 있다면 파악하는 것이 중요하다고 생각



월별로 주어진 Data 양의 차이가 큼

월 별 증감률 비교

< Target별 월에 따른 비율 변화 양상 >

3월부터 자동지급 비율이 증가하는 양상을 보이며
조사와 심사는 서로 Trade off 양상을 보임

2. 데이터 해석 2-2. 특성변수 EDA

MIRAE ASSET 미래에셋생명

자동지급 비율이 증가 추세를 보이는 것은, 판단 기준이 완화되었을 것이라고 추측가능.
실제로 Q&A 게시판을 통해 자동지급 기준과 관련된 다음과 같은 정보를 얻음.

[2019년 10월 29일 이후 변경된 자동지급의 분류 기준]
- 특수 조사 이력 고객에 대한 기준의 추가
- 실손 금액 기준 변경
- 질병 이력 관련 기준 삭제

이는 예측 데이터의 **바로 직전 달**인 11월부터
미래에셋 생명 내의 **분류 기준이 변경**됨을 뜻하는 중요한 정보

∴ 이러한 정보와 인사이트를 바탕으로 차후 모델링에 쓰일 데이터를 선택

LOVE AGE

3. 데이터 가공
(Feature Engineering)

3. 데이터 가공

MIRAE ASSET 미래에셋생명

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

3. 데이터 가공

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

3. 데이터 가공

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

3. 데이터 가공

기타 : **고객 관련 변수 생성**

< Target별 청구일-부활일 기간 코드 비율 >

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

LOVE AGE

4. 모델링

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

데이터 선택 근거 : EDA를 기반으로 데이터 기간에 대한 검증을 통해 활용 기간을 선택

< Target별 월에 따른 비율 변화 양상 >

[2019년 10월 29일 이후 변경된 자동지급의 분류 기준]

- 특수 조사 이력 고객에 대한 기준의 추가
- 실손 금액 기준 변경
- 질병 이력 관련 기준 삭제

p. 21

데이터 기간에 대한 검증의 필요성 제기

MIRAE ASSET 미래에셋생명

모델 선택 근거 : EDA 인사이트를 기반으로 모델 선택 근거와 기준을 고안함.

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

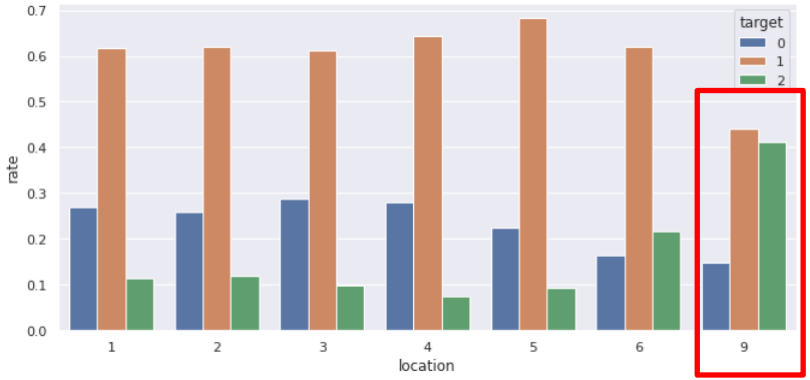
- Unknown 데이터를 모델에 반영하기 위해 구간을 나눠
  해당 데이터를 범위에서 벗어난 하나의 특성으로 반영해야 함 ex) 1, 2, 3, 9

MIRAE ASSET 미래에셋생명

모델 선택 근거

② Unknown 데이터를 제거하지 않고 모델에 반영

- Unknown 데이터가 다른 데이터에 비해 targe이 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)

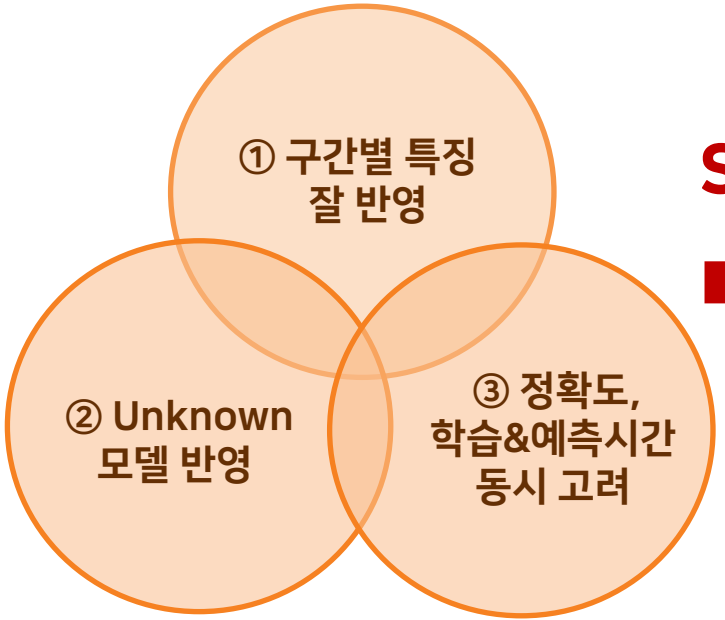
[No tuning]	학습시간(초)	예측시간(초)	F1_score
Logistic	8.54	0.008	0.418
Decision Tree	1.62	0.01	0.626
SVM	20270.68	0.0079	0.29

-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

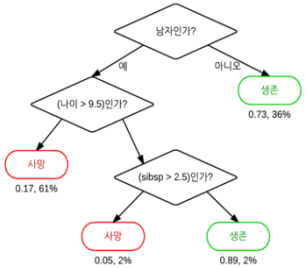
MIRAE ASSET 미래에셋생명

모델 계열 선택



Select

Decision Tree



- 지도학습에서 유용하게 사용 되고있는 기법 중 하나

- 학습에 사용되는 자료 집합을 적절하게 부분집합으로 나눔

- Decision Tree를 활용한 앙상블 모델들이 강력한 성능을 보임

https://ko.wikipedia.org/wiki/%EA%B2%B0%EC%A0%95_%ED%8A%B8%EB%A6%AC_%ED%95%99%EC%8A%B5%EB%B2%95

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)

MIRAE ASSET 미래에셋생명

검증 목표

√ 직전 달만 사용하여 target을 예측하는 것이 성능향상에 도움이 되는가?

√ 최적의 Ensemble algorithm은 무엇인가?

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

4. 모델링 4-3. 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

최종 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

1개월 이전의 데이터를 사용함으로 더 빠르고 좋은 성능을 낼 수 있음이 증명됨

Extra Tree에서 높은 score + 빠른 학습 시간

예측 시간이 가장 느리긴 하지만 그 차이가 미미함

➔ 1달 전의 데이터와 Extra Tree를 사용하기로 결정

MIRAE ASSET 미래에셋생명

최종 모델 선택 : Extra Tree (Extremely randomize tree)

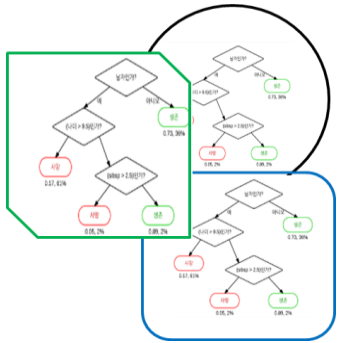
1) 알고리즘

√ 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



최종 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

MIRAE ASSET 미래에셋생명

모델 고도화 ① Extra tree의 랜덤성 보완

√ 모델의 랜덤성을 보완하기 위해 Extra tree에서 Random state를 다르게 하여 3가지 Random state의 결과를 반영함으로 안정성을 높임

√ 사례 기반 모델 추가 : KNN classifier

예측하고자 하는 target을 과거 경험(바로 전 달)에서 가장 유사한 data로 찾아내는 사례 기반 모델 (KNN)과 규칙기반모델의 (Tree) 결합으로 안정성을 높임

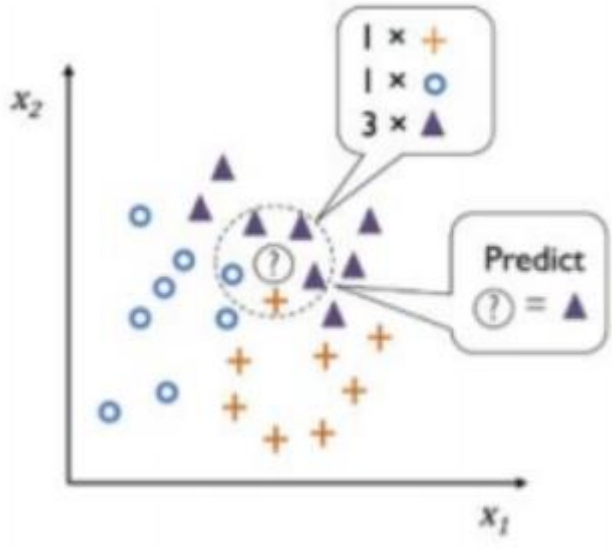
→ 한 모델의 편향을 학습을 다른 모델이 보완해주는 효과

→ 더욱 안정성을 높이고, 과적합을 방지

MIRAE ASSET 미래에셋생명

모델 고도화 ① Extra tree의 랜덤성 보완

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



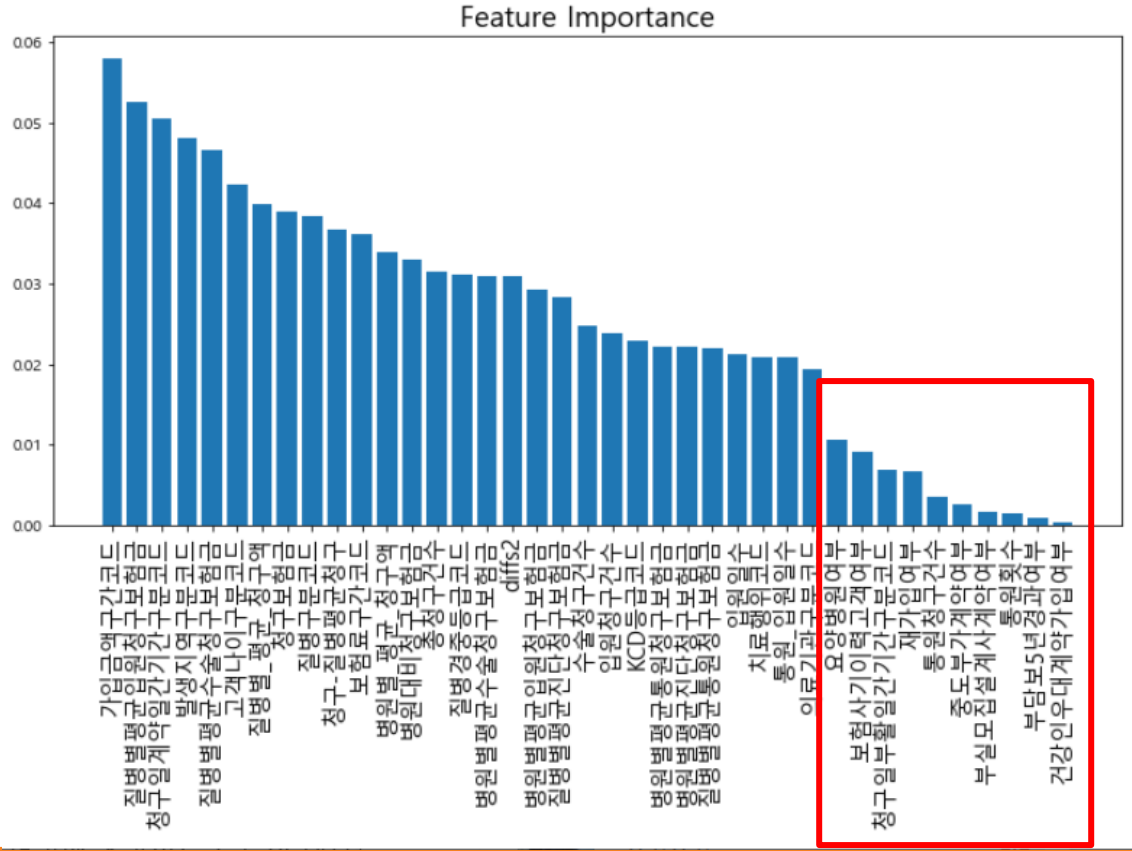
√ 기존 데이터와 가까운 이웃의 정보로
새로운 데이터를 예측하는 방법론

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

규칙기반 모델인 Extra Tree에 결합함으로써
전반적인 모델의 안정성을 높임

MIRAE ASSET 미래에셋생명

모델 고도화 ② Feature Selection



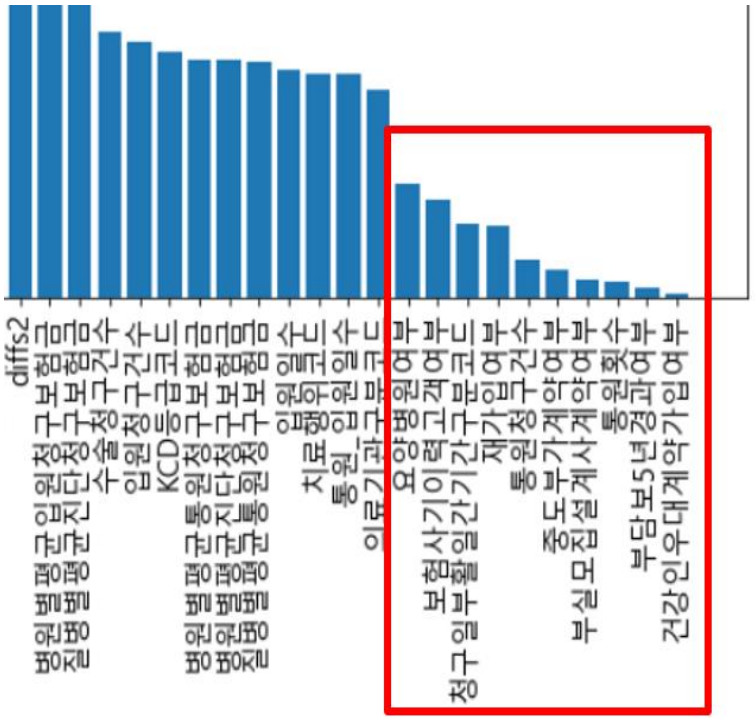
"Feature Importance"

해당 feature가 model에 적용되는 과정에서 전체적으로 classification의 불순도 (Gini index)를 얼마나 감소시키는지 나타냄

$$G.I(A) = \sum_{i=1}^{d} \left(R_i \left(1 - \sum_{k=1}^{m} p_{ik}^2 \right) \right)$$

MIRAE ASSET 미래에셋생명

모델 고도화 ② Feature Selection



√ 지니 정보이득이 급격히 감소한 Feature부터 Selection 진행
(Feature Importance에서 급격히 감소한 부분)

√ 이런 변수들이 오히려 학습을 방해할 수도 있기 때문에 F1 score를 고려하면서 제거

√ Feature를 줄임으로 차원의 저주 완화

√ Selection 결과 다음과 같은 변수를 제거

'요양병원 여부', '중도 부가계약 여부', '건강인우대계약 가입 여부 '

LOVE AGE

5. 결론 및 제안

5-1. 최종 모델의 장점과 한계
5-2. 모델 개선방향 제안
5-3. 관련 비즈니스 시스템 제안

5. 결론 및 제안 5-1. 최종 모델 장점과 한계

MIRAE ASSET 미래에셋생명

최종 모델 장점

QnA 정보와 EDA, 기간별 성능비교로 바로 전 달 데이터를 활용하기로 결정하였고
바로 전 달 데이터를 사용하여 F1 Score와 학습 속도와 예측 시간을 대폭 향상시킴

Unknown Data를 하나의 특성으로 잘 인식할 수 있고
IF else에 적합한 Decision Tree 기반의 Model 고려

Tree based model인 Light GBM, Random Forest, Extra Tree를
비교한 결과 Extra Tree가 F1 Score와 학습 속도에서 우수하였음

5. 결론 및 제안 5-1. 최종 모델 장점과 한계

사례 기반 모델과 규칙 기반 모델의 결합으로 안정성을 높여 과적합 방지

- 3가지 Random state의 결과를 반영함으로 안정성을 높인 Extra Tree와 사례 기반 모델인 KNN Classifier를 결합

- 앞서 해당 데이터셋에서 규칙 기반 모델의 강점을 보였고, 이에 따라 Extra Tree에 가중치를 더 부여한 가중평균을 진행

바로 전 달 Data + 3 Random state Extra tree + KNN

Public Score 88.059

5. 결론 및 제안 5-1. 최종 모델의 장점과 한계

MIRAE ASSET 미래에셋생명

한계

① 비식별화 데이터

- 공모전을 진행하며 미래에셋생명의 보험 상품 약관, Q&A 게시판을 통해 보험 상품과 약관 등의 특수 정보를 많이 수집함. 하지만 이들은 비식별화 데이터였기에 활용하지 못하였음

- 식별화된 고객정보를 알 수 있다면 활용가능한 파생변수들이 많았을 거라 예상

Q. 암보험 가입 후 60일이 지난 후 암으로 판정되었습니다. 보험의 혜택을 받을 수 있습니까?

A. 문의에 대한 답변

혜택을 받을 수 없습니다. 각종 암에 대한 보장은 계약일 또는 부활일로 부터 90일이 경과한 다음 날부터 그 효력이 발생되기 때문입니다. 암관련 보험에 가입하신 분들은 실효되어 보험을 부활하는 일이 없도록 주의하십시요. 부활하는 경우에도 부활 후 90일 이 경과한 다음날부터 암에 대한 혜택이 시작되니 주의하시기 바랍니다."

(출처 : 미래에셋 생명 FAQ)

[Ex1]

각종 암에 대한 보장은 계약일/부활일로부터 90일이 지난 다음날부터 효력이 발생함

이를 활용해 암 판정 환자가 계약일 혹은 부활일로부터 90일이 지났는지 여부를 모델에 반영했다면 분류에서 좋은 기대효과 예상

5. 결론 및 제안 5-1. 최종 모델의 장점과 한계

Q. 입원비 지급한도가 있나요?

A. 문의에 대한 답변

네, 있습니다. 입원비의 지급일수는 1회 입원당 120일을 최고 한도로 합니다. 다만, 동일한 질병 또는 재해로 인하여 입원을 2회 이상 한 경우에는 1회 입원으로 보고서 각 입원일수를 합산하여 상기 규정을 준용합니다. 그러나 동일한 질병 또는 재해에 의한 입원이라도 입원급여금이 지급된 최종입원의 퇴원일부터 180일을 경과하여 개시한 입원은 새로운 입원으로 봅니다. ※ 판매상품 및 판매시기에 따라 약관의 보장내용이 달라질수 있으니 가입하신 보험의 계약일자에 해당하는 약관을 참고하시기 바랍니다."

(출처 : 미래에셋 생명 FAQ)

[Ex2]

입원비 지급일수는 1회 입원당 120일을 최고 한도로 함. 이를 활용해 120일을 기준으로 target에 대한 입원일수의 반영을 더 보완할 수 있었을 것

[Ex3]

- 보험은 과거에 발생했던 개인의 사고 발생 통계에 따라 위험률을 산정하고, 그 위험률에 따라 보험료가 책정됨 (개인의 과거 가족력, 청구 내역 등의 반영)

- 식별화 데이터로서 개별 위험 발생 가능성을 지표화할 수 있었다면, 보험 가입자의 보험료 = 개별적 위험 발생 가능성 * 보험금(렉서스의 법칙) 등을 활용한 변수를 생성할 수 있었을 것

MIRAE ASSET 미래에셋생명

한계

② 자동지급의 정밀도

- 자동지급과 심사, 조사를 모두 고려한 평균 F1 score를 위주로 보았다는 한계

√ 실제로 자동지급인 고객에게 모델이 자동지급이라고 예측하는 것 보다 지급 대상이 아닌 고객에게 모델이 자동지급이라고 예측해 보험금을 지급하는 경우가 적어야 될 것임

√ 이를 반영할 수 있는 지표는 **정밀도(Precision)**로 정밀도를 고려한 모델 구축이 비즈니스 활용에 더 많은 가치를 창출할 것이라고 판단

$$(Precision) = \frac{TP}{TP + FP}$$

		분류 결과	
		False	True
실제 정답	False	True Negative	False Positive
	True	False Negative	True Positive

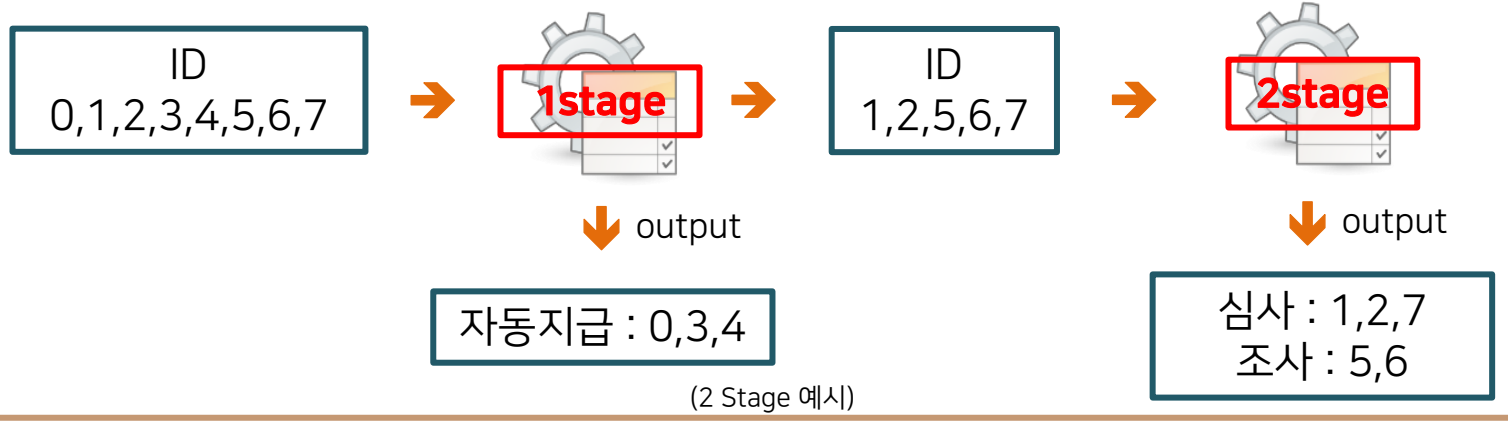
5. 결론 및 제안 5-2. 모델 개선 방향 제안

MIRAE ASSET 미래에셋생명

Two Stage model 제안

1. 배경 평균 F1 score 을 평가지표로 모델을 최적화 시켰기에, 자동지급의 정밀도가 낮아지는 한계 극복을 위해 이에 대한 분류를 강화할 수 있는 모델 필요

2. 내용 심사와 조사를 한 class로 놓고 자동지급과 그 외를 분류하는 이진 분류기를 먼저 구축한 뒤, 그 후 심사와 조사를 분류하는 2stage model



(2 Stage 예시)

5. 결론 및 제안 5-2. 모델 개선 방향 제안

Two Stage model 제안

3. 성능 검증 : 10월 데이터를 학습해 11월 데이터를 검증

2 Stage

실제 Target	예측 Target 0	예측 Target 1	예측 Target 2
0	7583	1696	30
1	2137	18217	235
2	29	1043	3990

F1 score : 0.84308 | Precision : 0.77782

기존 모델

실제 Target	예측 Target 0	예측 Target 1	예측 Target 2
0	7717	1566	26
1	2312	18024	253
2	55	994	4013

F1 score : 0.84313 | Precision : 0.76527

F1 Score도 고려하면서 **자동 지급에 대한 Precision을 강화**

5. 결론 및 제안 5-3. 관련 비즈니스 시스템 제안

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

5. 결론 및 제안 5-3. 관련 비즈니스 시스템 제안

MIRAE ASSET 미래에셋생명

4. 프로세스 식별화 데이터를 활용해 고객 군집을 생성

식별화 된 고객 데이터

-고객 건강관련 정보
-고객 신상 정보
-가족력 정보 등

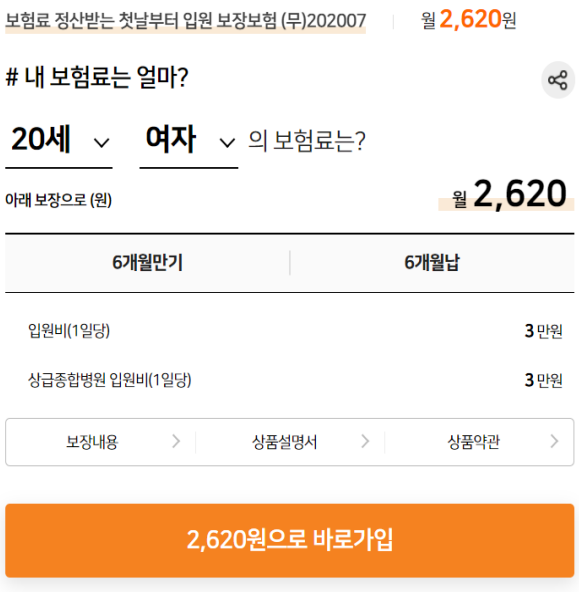
고객 클러스터링 모델 구축

고객을 군집으로 묶음
(Kmeans 등 비지도학습 사용)

[결과]

- 고객의 군집 정보를 우리 모델 데이터의 한 특성변수로 삽입 → 기존 모델의 성능 개선
- 군집별 사후분석 + 이상치 탐지 → 보험사기 예측의 지표로서 활용
- 유사 고객 군집 정보와 기존 온라인 서비스 연계 → 군집을 기반으로 커스터마이즈 된 보험상품의 추천

5. 결론 및 제안 5-3. 관련 비즈니스 시스템 제안



미래에셋생명 온라인 보험에 구축되어 있는
기존 보험료 계산 서비스

5. 기대효과

√ 식별화 정보를 기반으로 한 군집정보를 추가함으로써
기존 모델의 성능 향상 기대

√ 군집별 사후 분석을 통한 이상치 탐지를 통해
보험 사기 고객의 동향 파악
→ 보험 사기 예측의 지표로서 활용 가능
보험 사기 피해에 대비

√ 기존의 온라인 보험료 추천 서비스와 연계해
조금만 세부적인 정보를 입력하면
유사 고객이 가입한 보험상품이나 보험료 등을 제공
→ 개인화(customize)된 서비스 제공으로
고객 만족도 증대

6. 부록

6. 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