

2020 금융 빅데이터 페스티벌

미래에셋 머신러닝 경진대회 보험금 청구 건 분류

보험나라코딩공주 팀
최홍혁 서정민 이해원



팀 소개



성명 : 최홍혁
학교 : 고려대학교
학년 : 4학년

학과 : 통계학과
전화번호 : 010-2909-2056



성명 : 서정민
학교 : 송실대학교
학년 : 3학년

학과 : 산업정보시스템공학과
전화번호 : 010-9459-3906



성명 : 이혜원
학교 : 고려대학교
학년 : 석사과정

학과 : 통계학과
전화번호 : 010-4491-9411

CONTENTS

01 서론

- 문제 상황 및 배경 파악
- 주제 이해

02 데이터 해석

- 데이터 이해
- EDA

03 모델링

- Feature Engineering
- Model tuning & Evaluation

04 결론

- 비즈니스 활용 방안

PART

1

서론

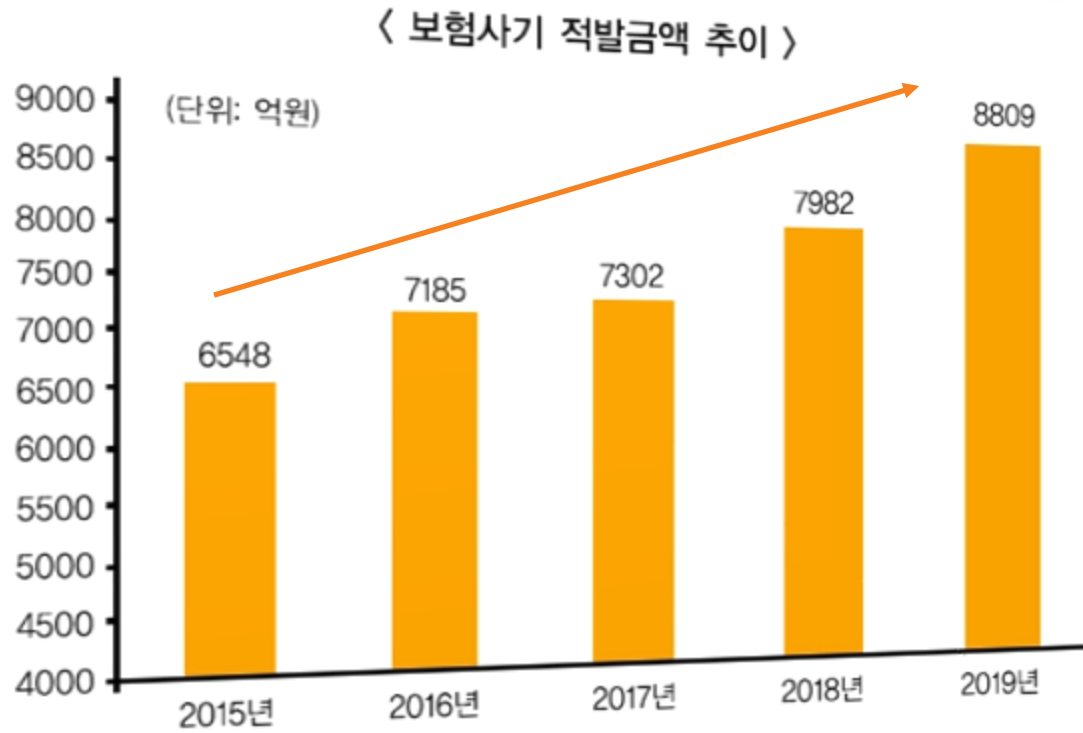
데이터 해석

모델링

결론

문제상황 및 배경 파악

보험 사기의 증가



* 자료: 금융감독원

- 보험사기 적발금액은 꾸준히 증가하여 작년 8809억원으로 최대치를 갱신했으며, 적발인원은 92538명으로 전년 대비 16.9% 상승했다.
- 일 평균 24억원, 254명의 보험사기가 적발되고 있다.

문제상황 및 배경 파악

보험 업계의 AI 사용

AI로 '더 빨라진' 보험업계...보험사기까지 잡아낸다

[보험사 AI 전성시대]①.보험업계, 계약 심사부터 업무 전자처리까지 AI 도입 활발

AI 도입 금융사, 시장 점유율 높아져...인력 채용 더 늘렸다

한화생명 '보험금 AI 자동심사 시스템' 특허청 기술특허 획득

현재 여러 국내 보험사들은 보험사기 예방을 위해 AI 시스템을 개발하고 실사용 중에 있다.
이러한 동향에 발맞추어 보험금 청구 분류를 예측하고 이를 보험사기 예방에 어떻게 활용할 수 있을지 탐구한다.

주제 이해

보험금 청구 프로세스

서론

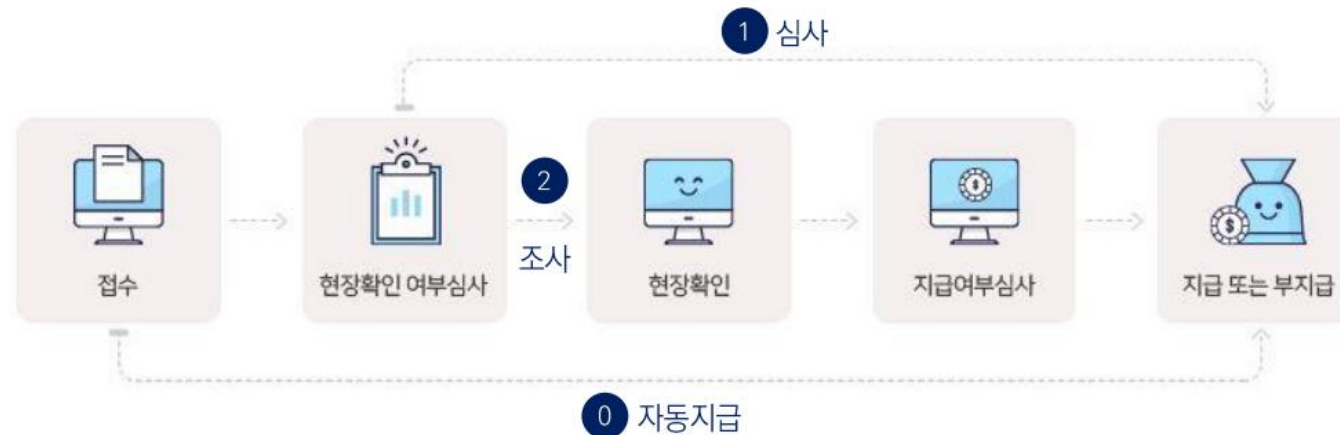
데이터 해석

모델링

결론

보험금 청구 프로세스

(자료: 당사 홈페이지)



2019년 1~11월까지의 월별 보험금 청구 데이터를 가지고
2019년 12월의 청구 건에 대한 분류 결과(자동지급, 심사, 조사)를 예측해야 한다.

주제 이해

보험금 청구 분류

서론

데이터 해석

모델링

결론

자동지급

접수 이후 별도의 조사 없이 보험금 지급 여부 결정.

심사

현장확인 여부심사 후 제출한 문서만으로 보험금 지급 여부를 판단 가능한 경우.

현장조사 필요 없음.

조사

제출한 서류만 가지고 판단하기 어렵거나, 보험 가입 시 보험사에 알려야 할 진료기록, 병력, 투약기록 등이 제대로 알려졌는지 확인이 필요한 경우.

현장조사 필요.

주제 이해

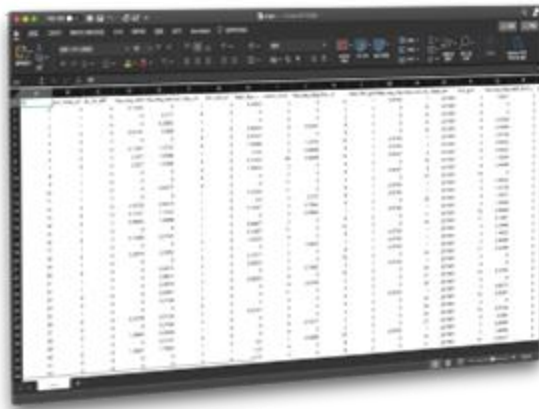
대회 목표

서론

데이터 해석

모델링

결론

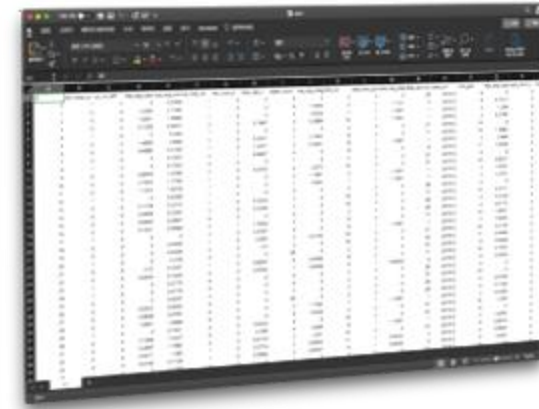
A screenshot of a data table with multiple columns and rows, representing the training data set. The table is displayed in a software interface with a dark theme.

<train data set>

2019년 1~11월까지 377928개의 보험금 청구 데이터
질병정보, 고객정보, 판매자정보 등 34개 변수



AI 모델 구축

A screenshot of a data table with multiple columns and rows, representing the test data set. The table is displayed in a software interface with a dark theme.

<test data set>

2019년 12월 22072개의 보험금 청구 데이터
자동지급 / 심사 / 조사 분류 예측

PART 2

서론

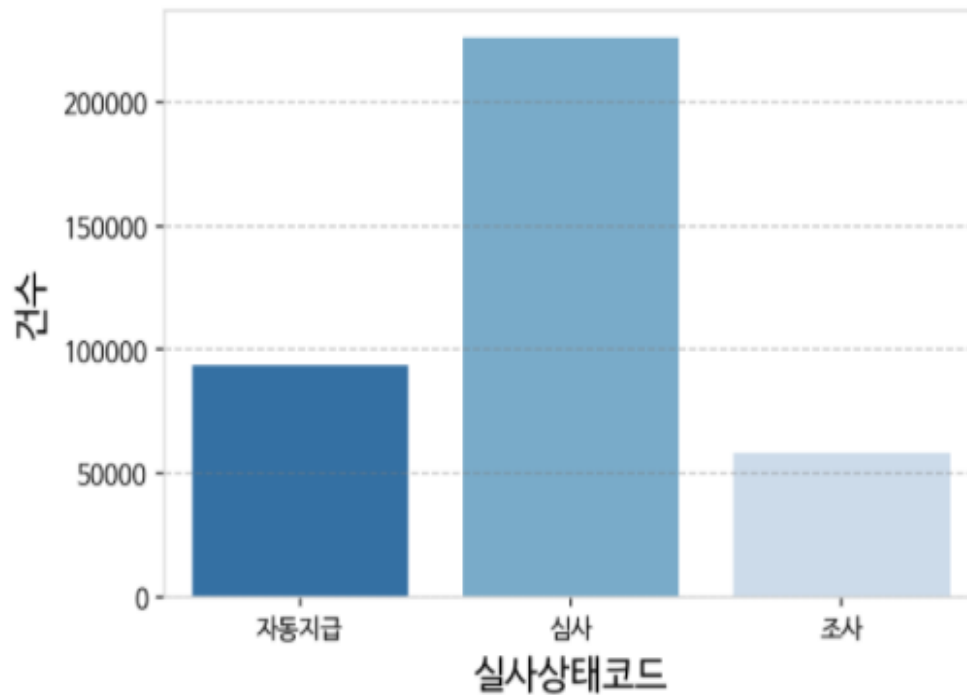
데이터 해석

모델링

결론

데이터 이해

보험금 청구 건 분류(target) 분포



심사(1)>자동지급(0)>조사(2) 순으로 데이터 존재
→ 타겟 분포가 불균형하다.

데이터 이해

데이터 변수 탐색

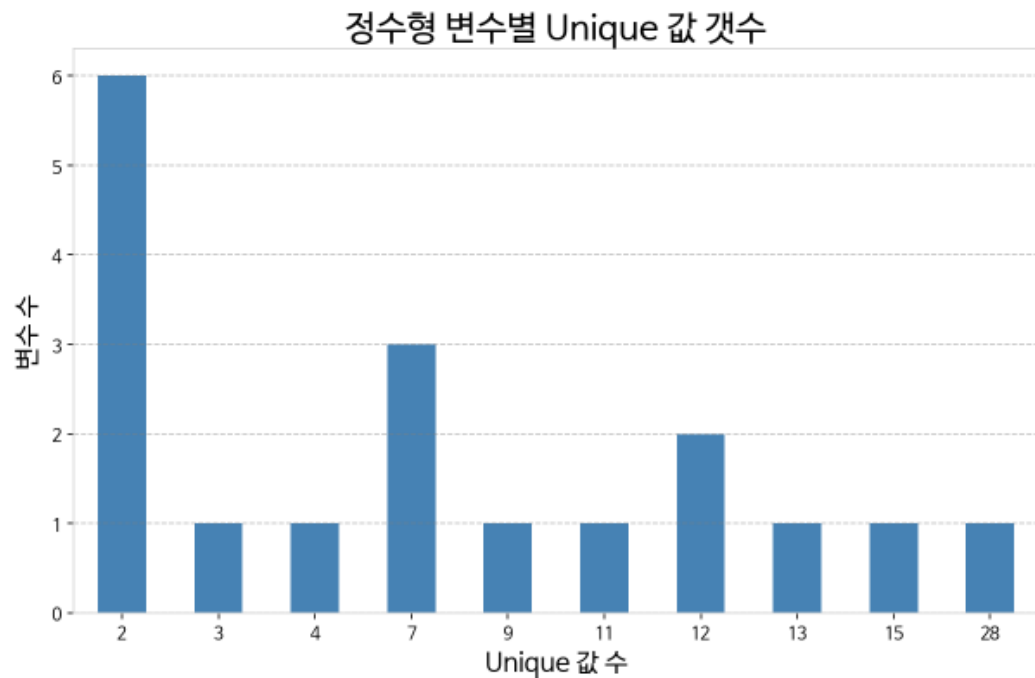
서론

데이터 해석

모델링

결론

정수형 변수

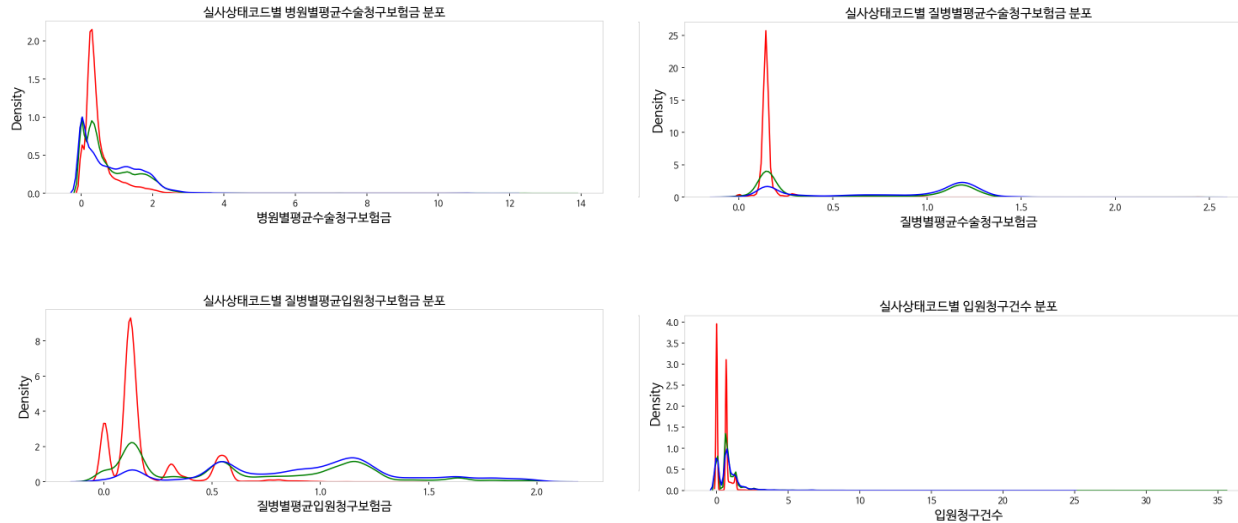


접수일련번호, Target 을 제외하면 정수형 변수는 총 18개이다.
Binary(Boolean type)변수가 6개로 가장 많으며,
categorical 변수는 12개로 다양한 범주를 가지고 있다.

데이터 이해

데이터 변수 탐색

실수형 변수



참고 빨간색 - 자동지급(0) / 초록색 - 심사(1) / 파란색 - 조사(2)

결론

실수형 변수 값에 따라 Target 의 커널밀도함수가 상이하기 때문에, 보험금 청구건 분류 결과를 예측할 때 실수형 변수들은 중요한 변수로 작용한다고 예상할 수 있다.

실수형 변수 KDE plot 해석

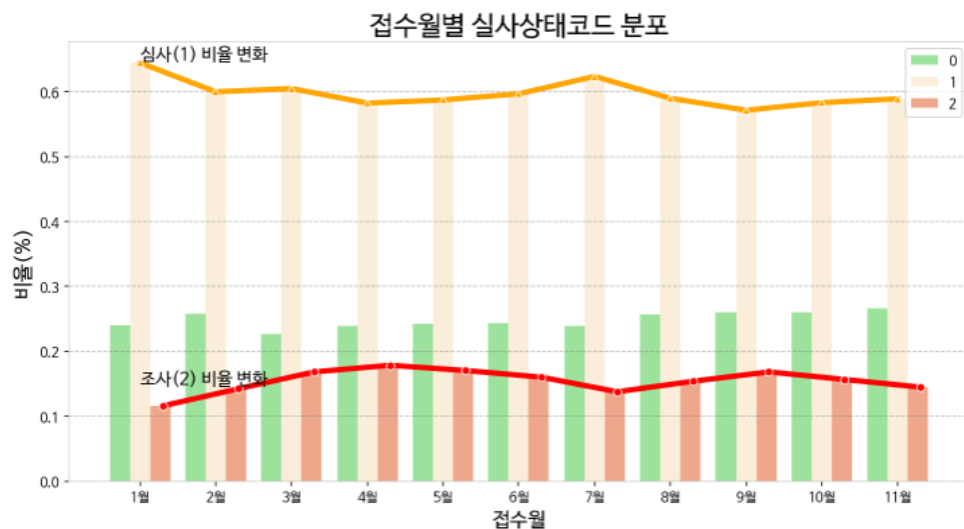
→ 각 실수형 변수의 값에 따라 보험금 청구건 분류 결과가 상이하게 나타난다.

예시 1) 질병별수술청구보험금이 0에 근접한 값을 가지면 대부분 자동지급(0)으로 분류된다.

EDA

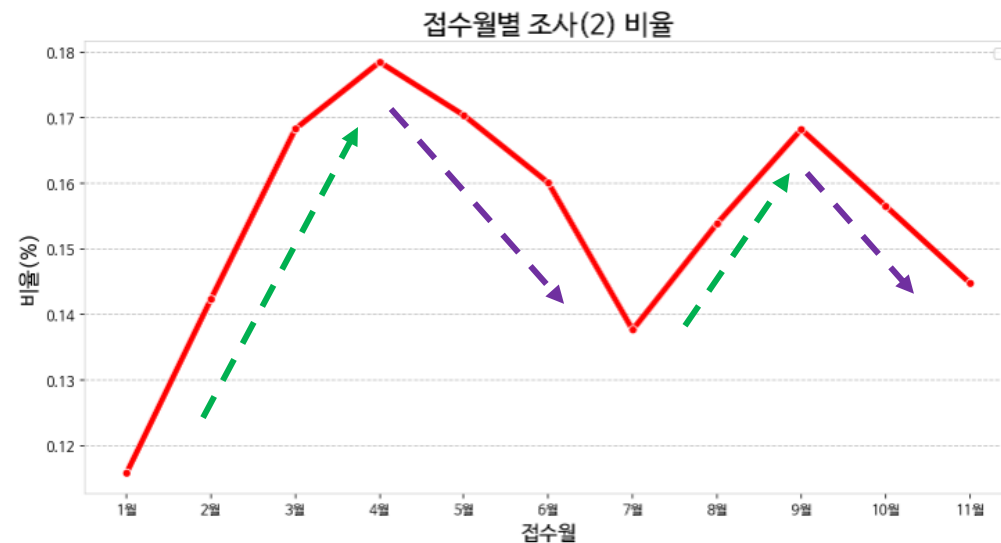
접수 월(month)

서론 데이터 해석 모델링 결론



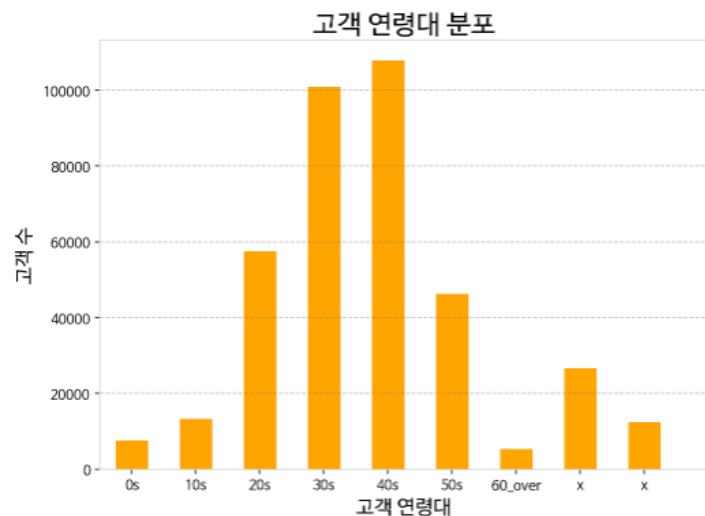
그래프 해석

- 조사(2)의 비율은 1월부터 4월까지 3배 가까이 상승했지만 4월부터 7월까지 하락한다. 그리고 9월까지 증가 후 다시 하락하는 패턴을 보인다.
- 조사(2)가 차지하는 비율을 보면, 특히 1월, 7월은 다른 월(month)에 비해 낮다.
- 심사(1)의 비율 변화는 조사(2)의 비율 변화와 반대되는 패턴을 보인다.



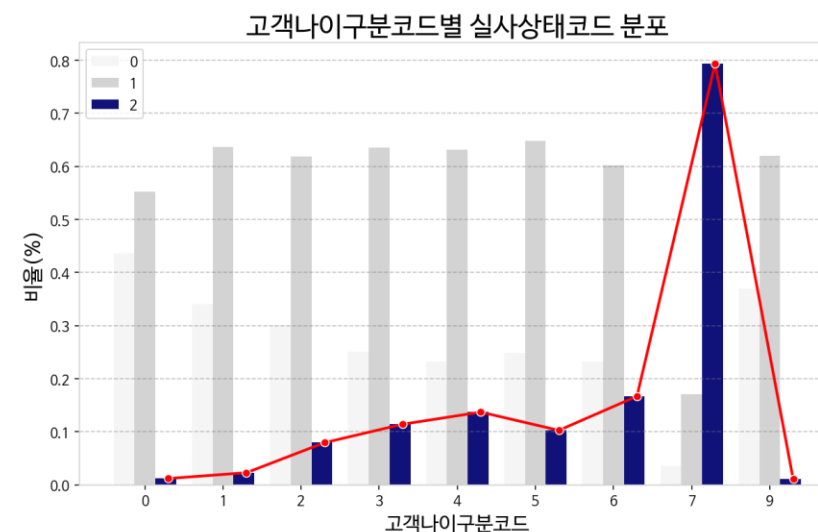
결론

- (1) 보험금 청구건 분류 결과에는 **계절성**이 중요하게 반영된다고 판단했다.
- (2) 조사 및 심사 비율 변화의 패턴을 분석하면 **12월에는 조사의 비율은 낮아지고 심사의 비율은 높아질 것으로 예상된다.**



보험금 청구 고객의 연령대의 분포 분석 결과

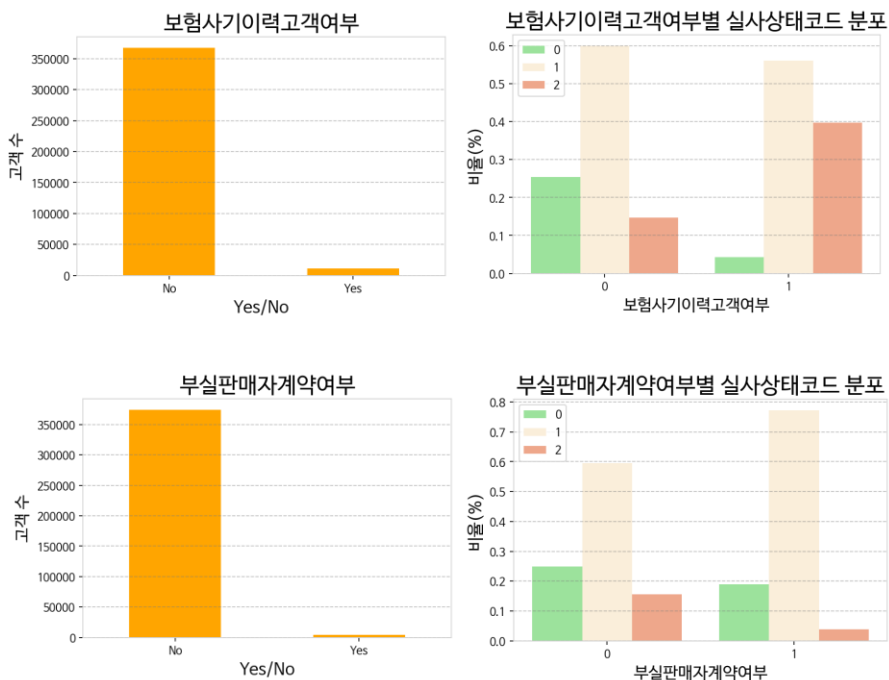
- 20대~50대 고객이 대부분을 차지한다.
- 60대 이상의 고객은 0대, 10대 고객보다도 보험금 청구 수가 적다.



대체적으로 연령대가 증가할수록 조사(2)가 차지하는 비율이 높아지는 경향을 보인다.

특이하게도, 고객나이구분코드가 7인 경우 조사(2)의 비율이 다른 나이구분코드에 비해 압도적으로 높다. 해당 내용은 뒷부분에서 자세히 다루겠다.

* 고객연령대분포 그래프에서 x로 표시한 항목은 고객나이구분코드 7, 9에 해당하는 것으로, 해당 코드로는 연령대 정보를 알 수 없기 때문에 x로 표시했습니다.



그래프 해석

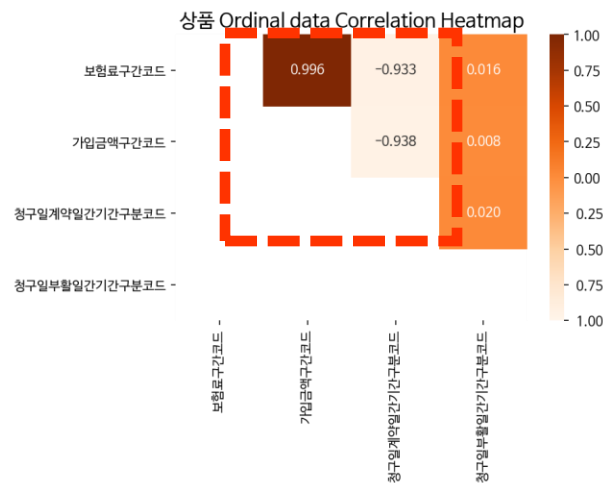
보험사기이력이 있거나 부실판매자와 계약한 고객의 비율은 매우 낮다. 그러나 두 상황에 해당하는 경우 그렇지 않을 때와 비교해 실사상태코드의 분포가 상이했다.

결론

보험금을 청구한 고객이 보험사기이력이 있는 경우 부정행위에 대한 위험성이 증가하여 해당 보험금 청구건을 조사(2)로 분류할 확률이 높아진다.

부실판매자계약여부는 binary (0, 1) 값마다 실사상태코드 분포가 상이했다.

상관계수 Heatmap 해석



<상관계수 히트맵>

- 보험료구간코드 & 가입금액구간코드의 상관계수: 0.996
- 보험료구간코드 & 청구일계약일기간구분코드의 상관계수: -0.933

구간코드 변수들은 매우 강한 상관관계를 보인다.

구간코드 99값 비교

	보험료구간코드	가입금액구간코드
7	99	99
50	99	99
96	99	99
105	99	99
108	99	99
...
377919	99	99
377920	99	99
377922	99	99
377924	99	99
377925	99	99

34436 rows × 2 columns

	보험료구간코드	가입금액구간코드	청구일계약일기간구분코드
7	99	99	0
50	99	99	0
96	99	99	0
105	99	99	0
108	99	99	0
...
377919	99	99	0
377920	99	99	0
377922	99	99	0
377924	99	99	0
377925	99	99	0

34436 rows × 3 columns

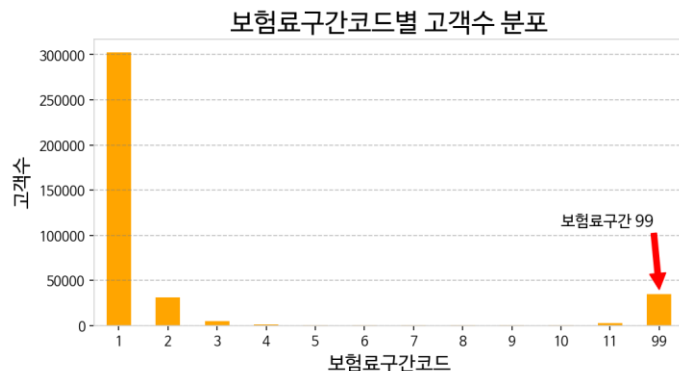
보험료구간코드가 99(Unknown)인 데이터와 가입금액구간코드가 99(Unknown)인 데이터를 추출결과

(1) 보험료구간코드 99(Unknown) → 가입금액구간코드 99(Unknown)

(2) 보험료구간코드, 가입금액구간코드가 99(Unknown) → 청구일계약일기간구분코드 0(Unknown)

이러한 이유로 구간코드 변수가 절댓값 0.9 이상의 강한 상관관계를 보인다.

(1)

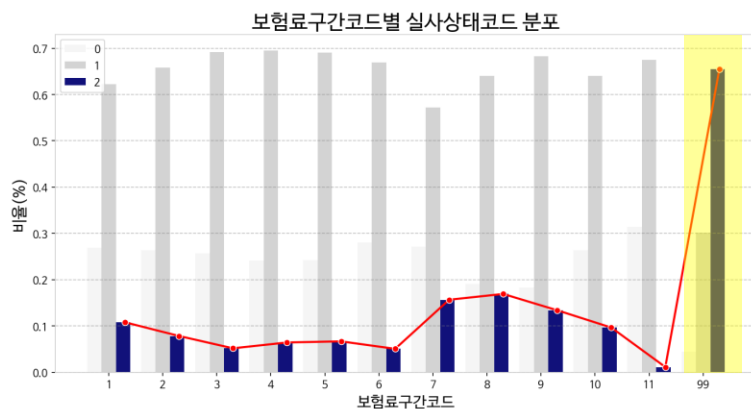


그래프 해석

1(0~10만원 미만)에 해당하는 고객수가 가장 많고, 그 다음으로는 99(Unkown)이 많다.

➡ 대부분의 고객이 10만원 미만의 소액의 보험료를 납입하는 경향이 있다고 볼 수 있다.

(2)



그래프 해석

특이하게도 보험료구간코드가 99이면 조사(2)가 차지하는 비율이 다른 구간코드보다 월등히 높다.

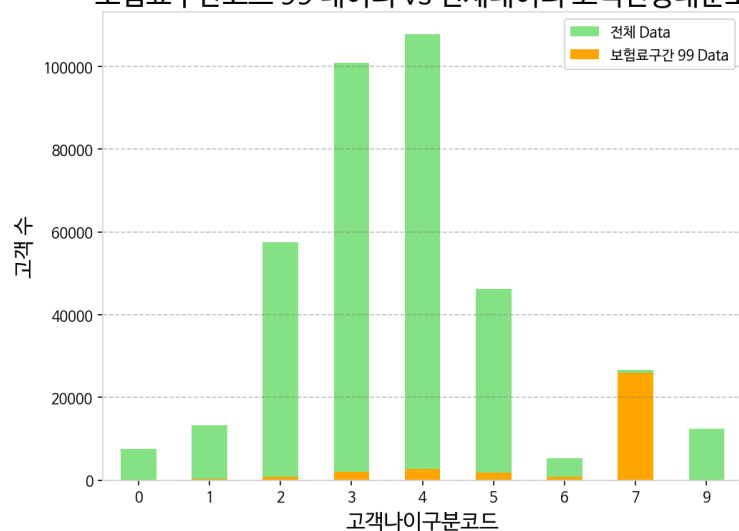
앞에서 고객데이터 분석을 실시했을 때, 고객나이구분코드가 7인 경우 조사(2)의 비율이 압도적으로 높아지는 특징을 발견할 수 있었다.

➤ 보험료구간코드 99(UnKnown) ↔ 고객나이구분코드 7(Unknown)의 관계성을 예상할 수 있다.

과정

1. 보험료구간코드 99에 해당하는 데이터만 추출하여 data_99 를 생성한 다음, data_99의 고객나이구분코드별 고객 수 분포를 확인
2. Data_99와 전체 데이터셋의 고객나이구분코드별 고객 수 분포 비교

보험료구간코드 99 데이터 vs 전체데이터 고객연령대분포



참고)
보험료구간 99 Data: 보험료구간코드 99에 해당하는 데이터만 추출

그래프 해석

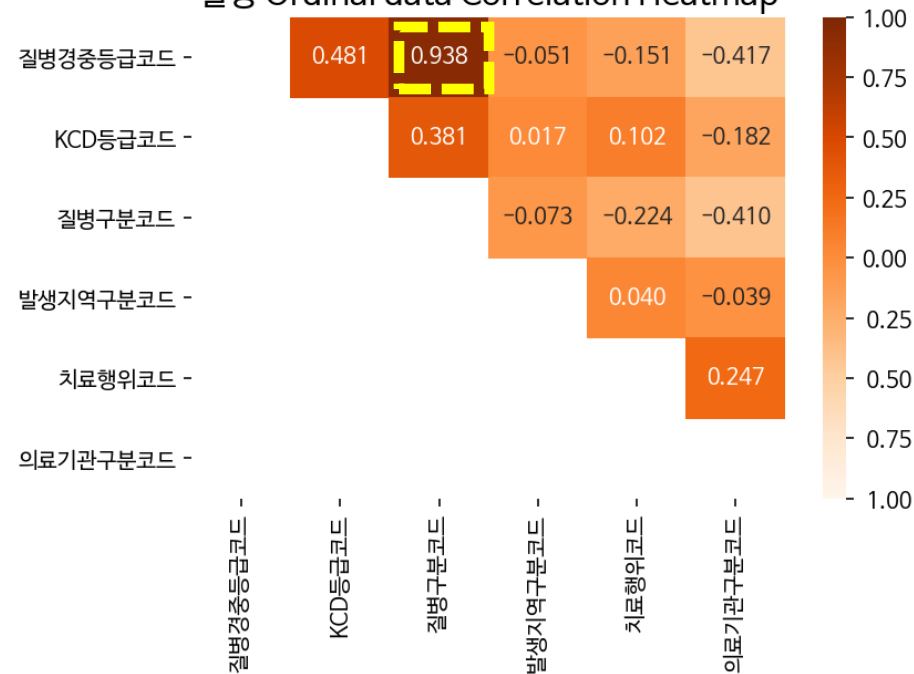
- data_99에 속하는 고객의 나이는 대부분 카테고리 7에 속한다.
- 전체 데이터와 비교했을 때, 전체 데이터에서 고객나이구분코드 7에 속하는 고객의 99%는 납입 보험료 구간이 Unkown(99)에 해당된다.

» 결론

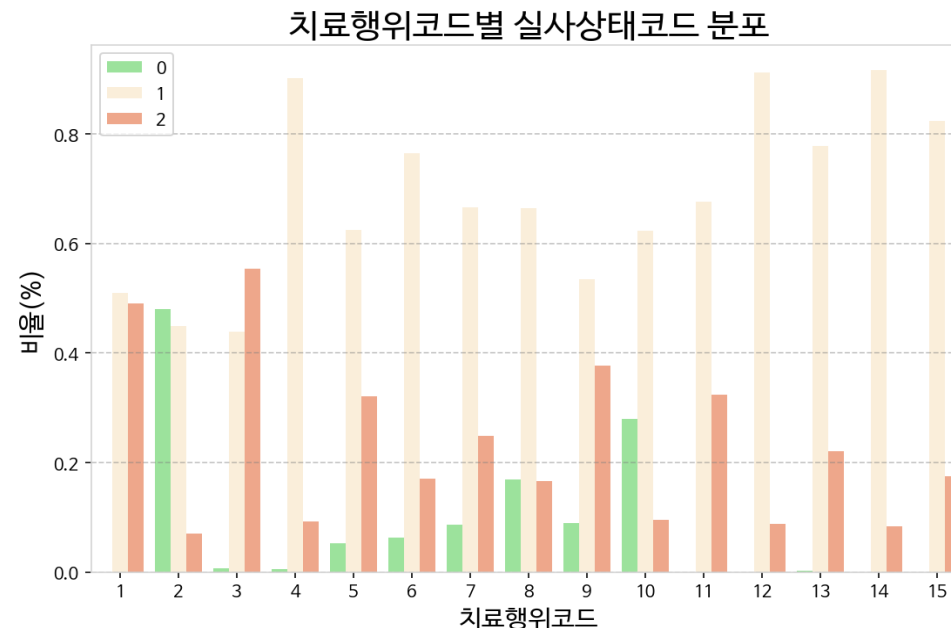
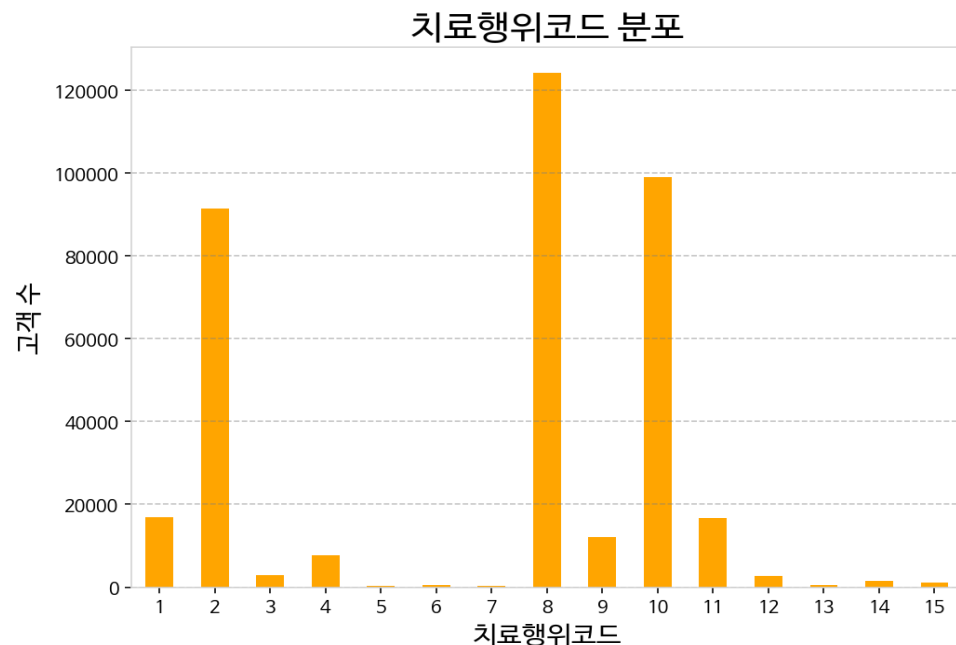
보험금 청구건 데이터에서 고객나이구분코드, 상품 ordinal 변수 중 하나라도 Unkown 값이 발생하면, 다른 변수들도 대부분 Unkown 값을 갖는다.

1	질병경중등급코드	2	질병구분코드
질병경중등급코드	질병경중등급명	질병구분코드	질병명
1	중증	1	암
		2	상피내암
		3	경계성
		4	심장질환
		5	
		6	뇌혈관질환
2	성인	7	간질환
		8	
		9	신장질환
		10	감상선질환
		11	폐렴
		12	천식
		13	위궤양
		14	십이지장궤양
3	생활	15	고혈압
		16	당뇨병
		17	관절염
		18	
		19	
		20	
		21	골다공증
		22	백내장
		23	중이염
		24	충수염
		25	남성비뇨기계
		26	
		27	부인과
		28	

질병 Ordinal data Correlation Heatmap



질병 데이터에 속하는 ordinal 변수들의 상관계수 Heatmap을 보면 질병경중등급코드와 질병구분코드의 상관계수는 0.938로 매우 높은 것을 확인할 수 있다. 이는 질병구분코드가 질병경중등급코드의 하위 카테고리로 나누어져 있기 때문인 것으로 보인다.



대부분의 보험금 청구 고객은 입원(2), 수술(8), 입원+수술(10) 치료를 받고 보험금을 청구했다. 치료행위코드별 실사상태코드 분포를 확인해본 결과, 진단행위가 포함될 경우 조사 비율이 높았다. 따라서 어떤 치료행위를 시행했냐에 따라서 실사상태코드의 분포가 달라진다고 판단되었다.

- 접수년월별로 실사상태코드의 분포가 변화하는 양상을 띄었기 때문에 계절적 요소를 추가해야 할 것으로 판단되었다.
- Binary type의 변수들 중에 조사의 비율에 영향을 미치는 변수들이 있으므로 이를 이용해 새로운 변수를 만들어 보고자 하였다.
- 고객나이구분코드, 보험료구간코드, 가입금액구간코드 및 청구일계약일간구분코드의 unknown 값은 서로 관련성이 높았다. 따라서 이러한 unknown여부를 나타내는 새로운 변수가 필요한 것으로 판단되었다.
- 질병치료행위별로 실사상태코드 분포가 다르게 나타났다. 따라서 질병치료행위(입원, 통원, 수술, 진단) 여부를 나타내는 새로운 변수를 만들어 보고자 하였다.

PART

3

서론

데이터 해석

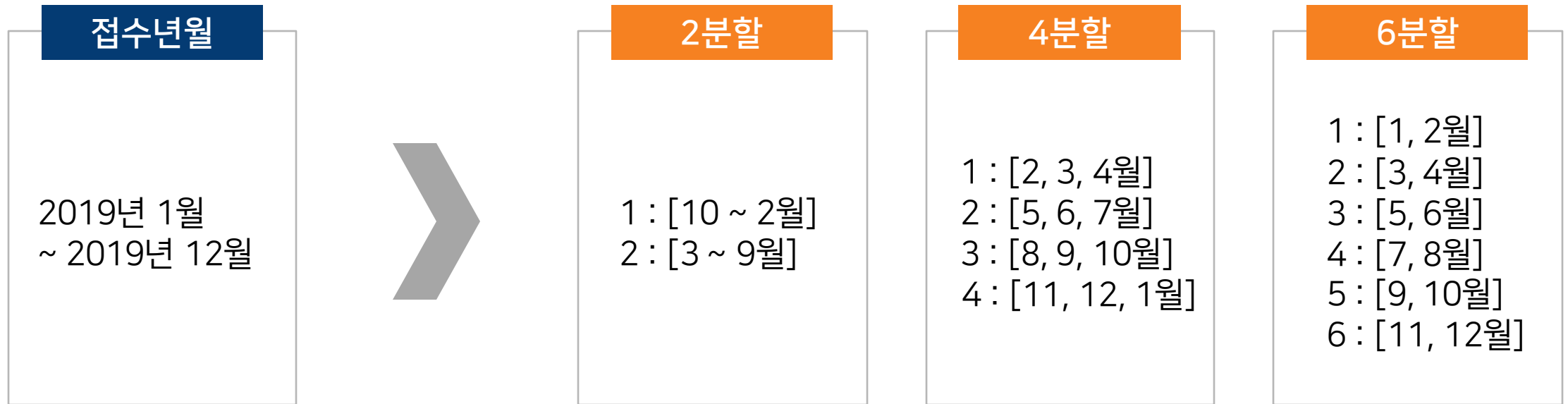
모델링

결론

Feature Engineering

변수 선택

계절성



접수년월 변수에서 월별로 grouping하여 계절, 6분할, 2분할의 세 변수 생성

Feature Engineering

변수 선택

의료기관구분코드



train data에는 의료기관구분코드의 카테고리가 1,2,3,9로 존재하나
test data에는 카테고리 9 데이터가 존재하지 않아 제거해 주었다.

서론

데이터 해석

모델링

결론

Feature Engineering

변수 선택

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

치료 행위

치료 행위 중 입원, 통원, 수술 여부를 Binary 변수(입원 여부, 통원여부, 수술여부)로 생성하였다.

진단여부는 추가시 오히려 성능이 하락하여 제거하였다.

치료행위여부 변수 추가 후 따로 치료행위코드 변수를 제거하지는 않았다.

Feature Engineering

변수 선택

서론

데이터 해석

모델링

결론

시도하였으나 성능 저하로 수정하지 않은 변수

Null값 여부

보험료구간코드와 가입금액구간코드 값의 99여부를 나타내는 binary 변수를 추가

부정점수

1 값을 가질 때 조사 비율이 높아지는 binary 변수(Null값 여부, 진단여부, 보험사 기이력고객여부)들의 합계를 부정점수 변수로 추가

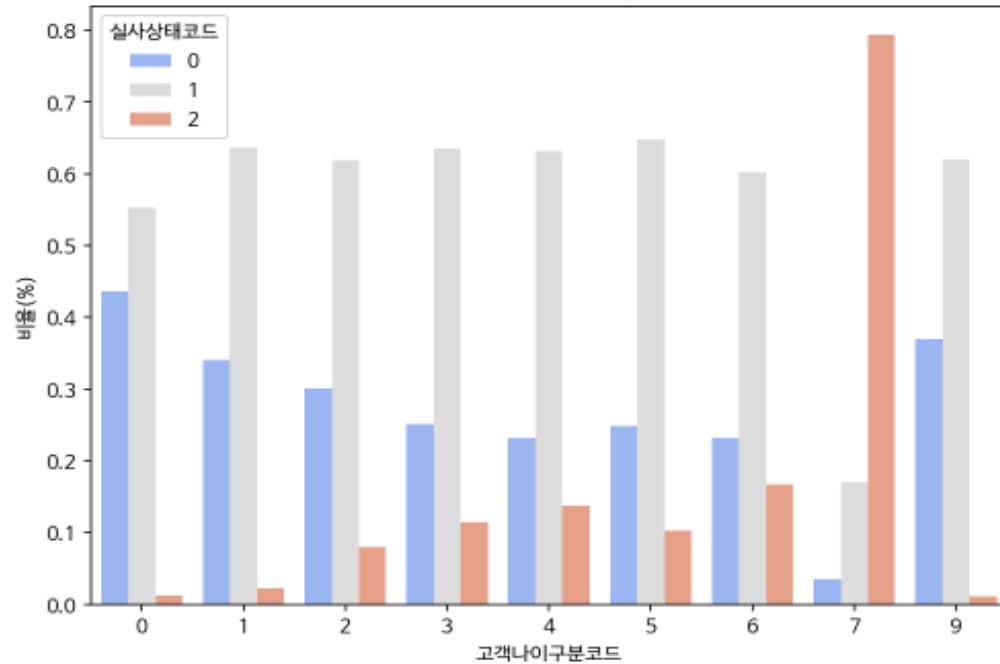
고객나이 구분코드

과제설명자료에 나와있지 않은 고객나이구분코드 카테고리(7, 9)를 수정

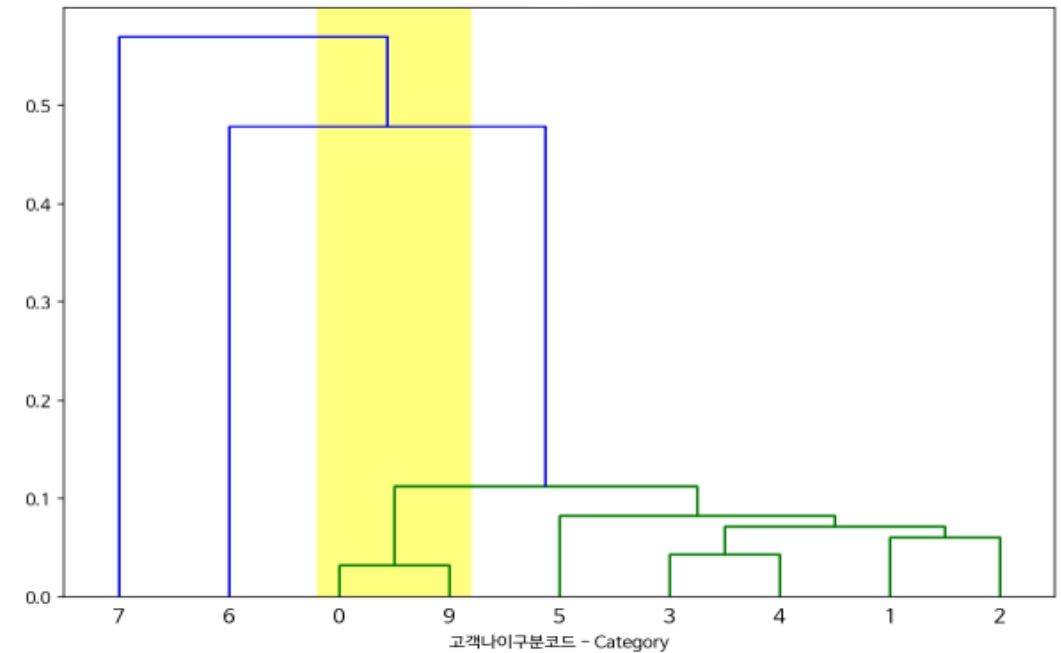
Feature Engineering

고객나이구분코드

나이별 실사상태코드 분포



고객나이구분코드 Dendrogram



나이별 실사상태코드 분포 및 Hierarchical Clustering 결과를 통해 카테고리 9은 0과 병합하고, 카테고리 7은 결측치로 판단하여 분석했으나 성능 저하로 원 자료를 사용하였다.

Model Tuning & Evaluation

모델 선택

Random Forest, XGBoost, CatBoost, LightGBM, RNN 등 많은 모델을 사용해 보았으나 최종적으로 LightGBM을 선택하였다.

LightGBM

XGBoost, CatBoost와 같은 Boosting 기반 모델은 단일 모델 대비 높은 분류 성능을 가진다. 특히 LightGBM은 XGBoost 대비 학습 시간이 짧아 많은 파라미터를 튜닝하기에 유리하다. 이번 보험 청구 데이터는 분석할 때 하이퍼 파라미터에 따라 예측 결과의 차이가 많이 나서 파라미터 튜닝에 유리한 LightGBM으로 모델을 선정하였다.

Model Tuning & Evaluation

하이퍼 파라미터 튜닝

서론

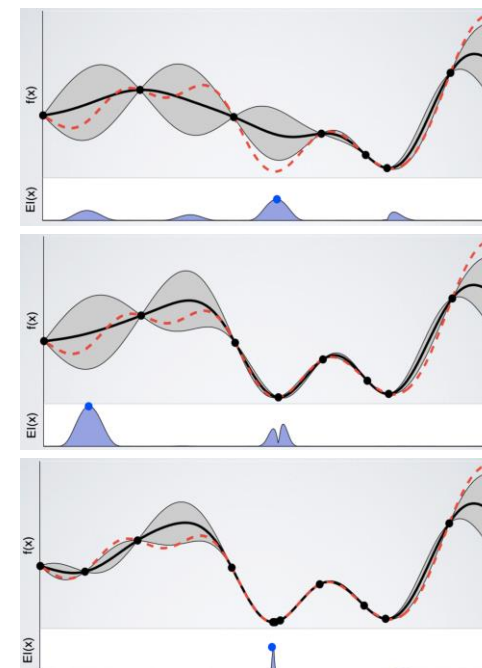
데이터 해석

모델링

결론

Bayesian Optimization

iter	target	colsam...	learni...	max_depth	min_ch...	min_ch...	n_esti...	num_le...
1	0.9579	0.8586	0.223	170.1	2.124	6.919	46.97	1.831e+0
2	0.9594	0.8315	0.1939	135.4	12.54	2.095	57.98	1.578e+0
3	0.9657	0.9923	0.3181	861.4	12.9	7.533	95.33	1.217e+0
4	0.8678	0.9915	0.1445	498.4	7.935	6.931	10.93	1.591e+0
5	0.9669	0.7174	0.3953	882.7	15.2	4.429	97.21	1.892e+0
6	0.9636	0.7461	0.3532	700.4	19.94	9.779	96.45	1.721e+0
7	0.875	0.7969	0.1032	297.0	5.616	0.8012	19.88	1.197e+0
8	0.7522	0.8249	0.03803	10.76	5.815	6.786	33.83	960.3
9	0.9585	0.7439	0.2774	820.9	15.29	8.652	86.95	882.2
10	0.8049	0.8028	0.04359	912.2	3.913	7.052	56.47	320.2
11	0.8781	0.9187	0.1912	332.2	13.67	0.9111	8.317	1.975e+0
12	0.8133	0.7085	0.07372	128.9	3.832	2.488	41.03	295.3
13	0.88	0.9414	0.2167	936.1	17.94	6.914	7.639	1.913e+0
14	0.9297	0.7429	0.3643	419.1	14.31	8.652	65.86	330.7
15	0.9385	0.7472	0.3995	183.5	13.84	8.233	45.65	528.2
16	0.8272	0.8017	0.3039	358.5	1.633	0.3277	3.63	1.072e+0
17	0.8444	0.9502	0.1818	75.09	12.15	7.648	10.39	863.0
18	0.9433	0.9657	0.177	485.3	17.04	7.047	62.88	840.1
19	0.9604	0.7965	0.3954	873.2	5.002	2.574	46.3	1.426e+0



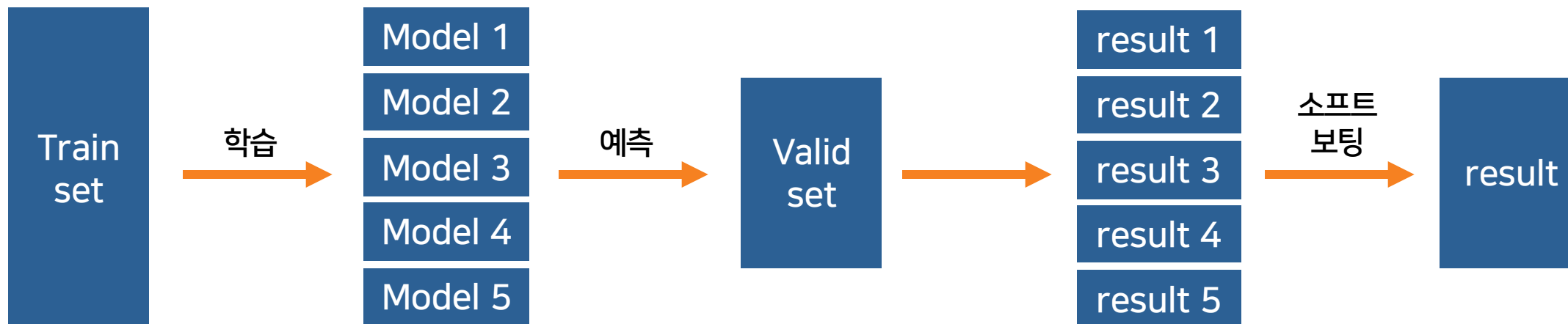
하이퍼 파라미터 튜닝을 위해 Bayesian Optimization을 사용했다.

Model Tuning & Evaluation

모델 설명

Step 1. 5개의 random seed를 이용해 LGBMClassifier 모델 5개를 생성

Step 2. LGBMClassifier의 predict_proba를 이용하여 5개 모델 결과를 소프트 보팅

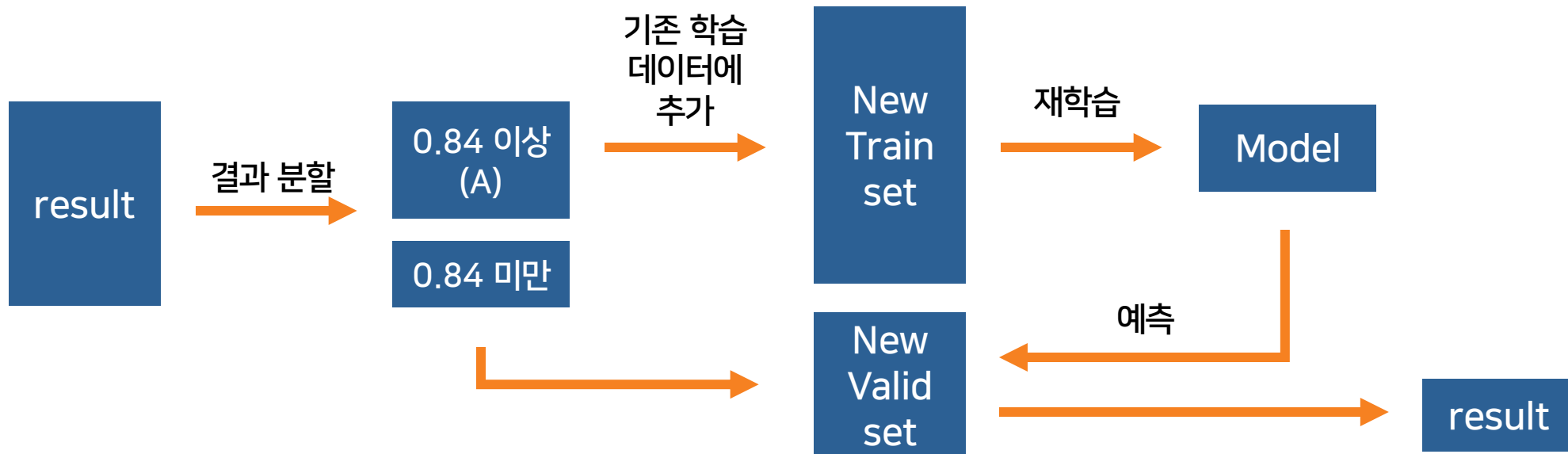


5개의 결과를 소프트 보팅함으로써 약 0.3~0.5점의 f1 score 상승을 얻을 수 있었다.

Model Tuning & Evaluation

모델 설명

Step 3. 소프트 보팅한 결과값에서 row별로 예측된 class들의 최대확률이 0.84 이상인 데이터는 예측이 맞았다고 가정하고 train set에 추가 후 남은 데이터만 valid set으로 재학습

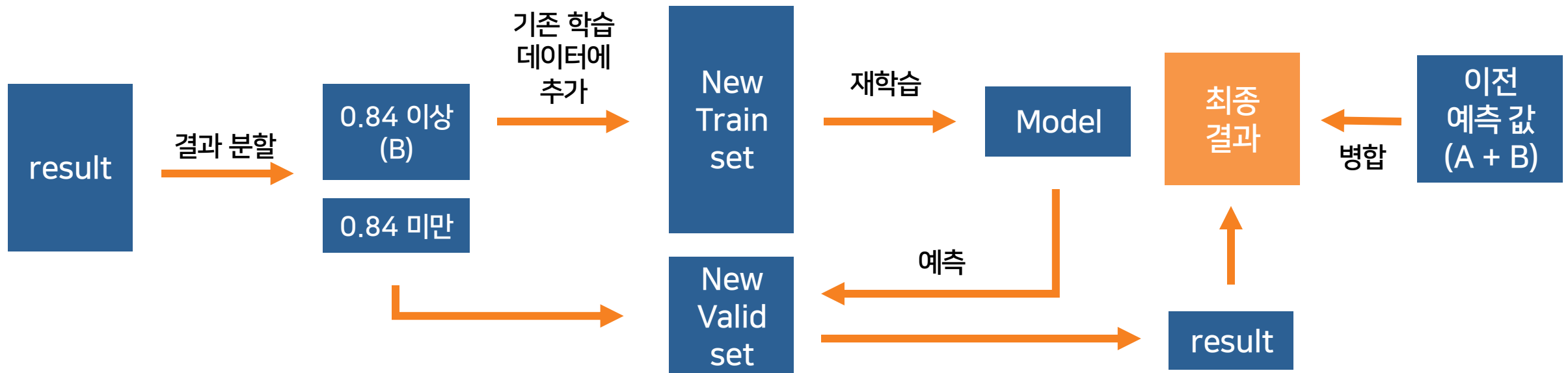


최대확률이 0.84 이상인 데이터를 학습 데이터에 추가해주면서 학습 데이터가 증가하는 효과가 나타나 예측 정확도가 상승했다.

Model Tuning & Evaluation

모델 설명

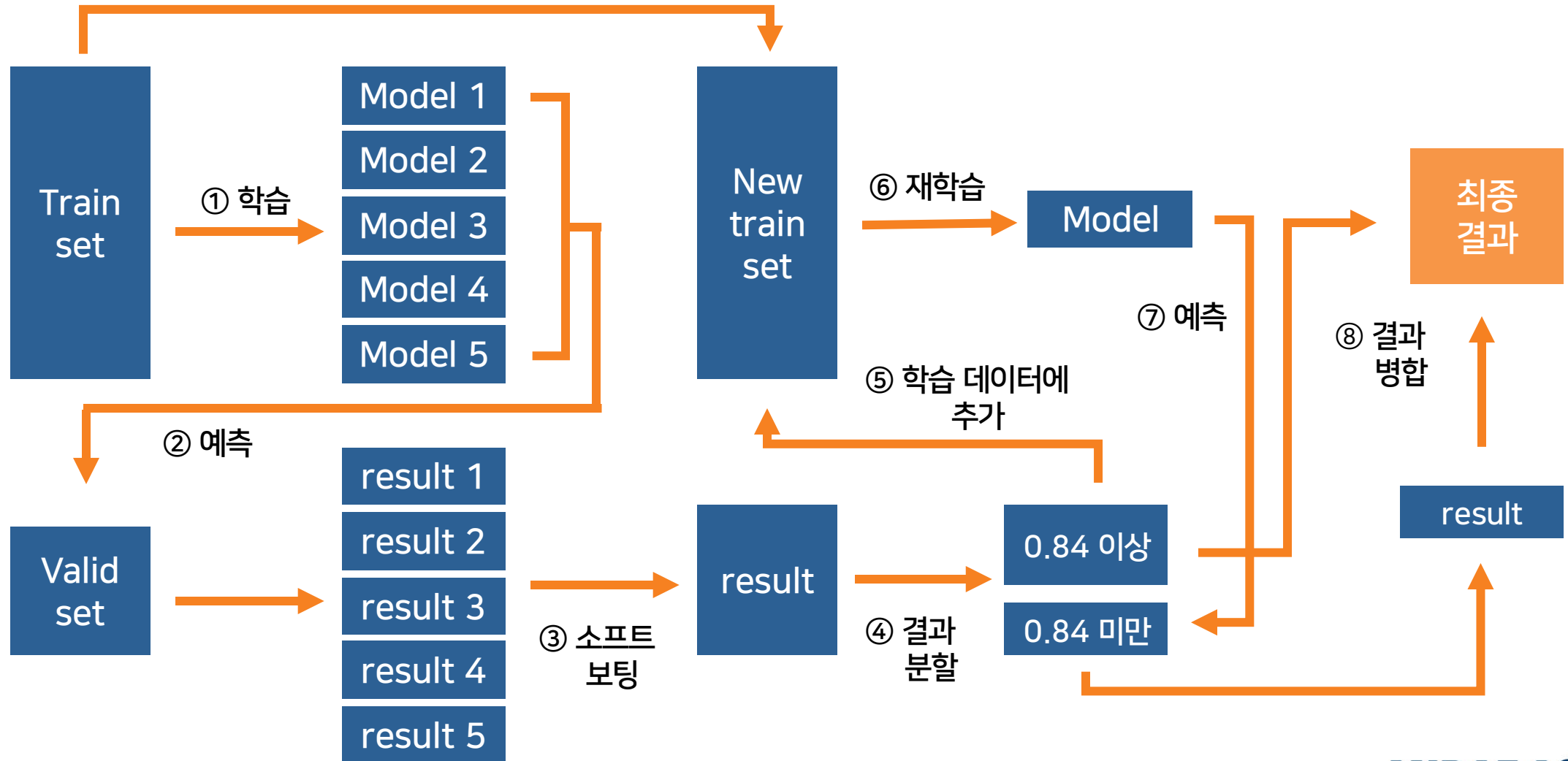
Step 4. 이전 과정(결과 값을 분할해서 학습 데이터에 추가하는 과정) 한번 더 반복 실행 후 초기 예측값과 병합



재학습 과정을 통해 약 0.5~0.8점의 f1 score 상승을 얻을 수 있었다.

Model Tuning & Evaluation

모델 구조화



Model Tuning & Evaluation

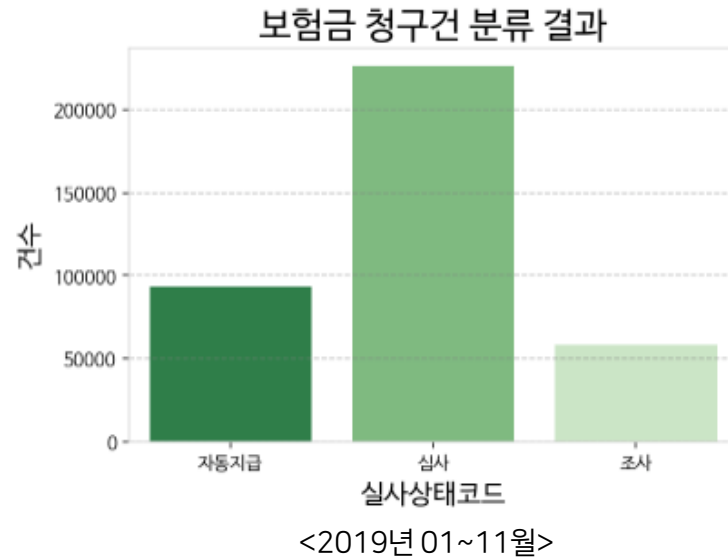
모델 평가

서론

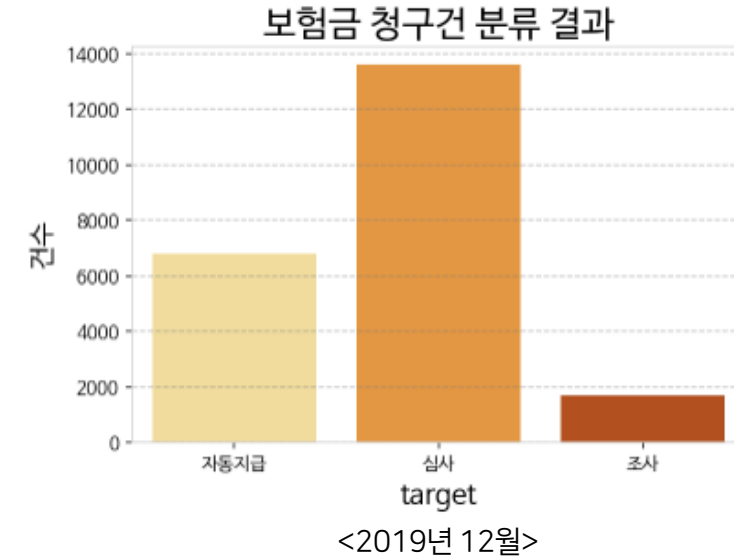
데이터 해석

모델링

결론



예측



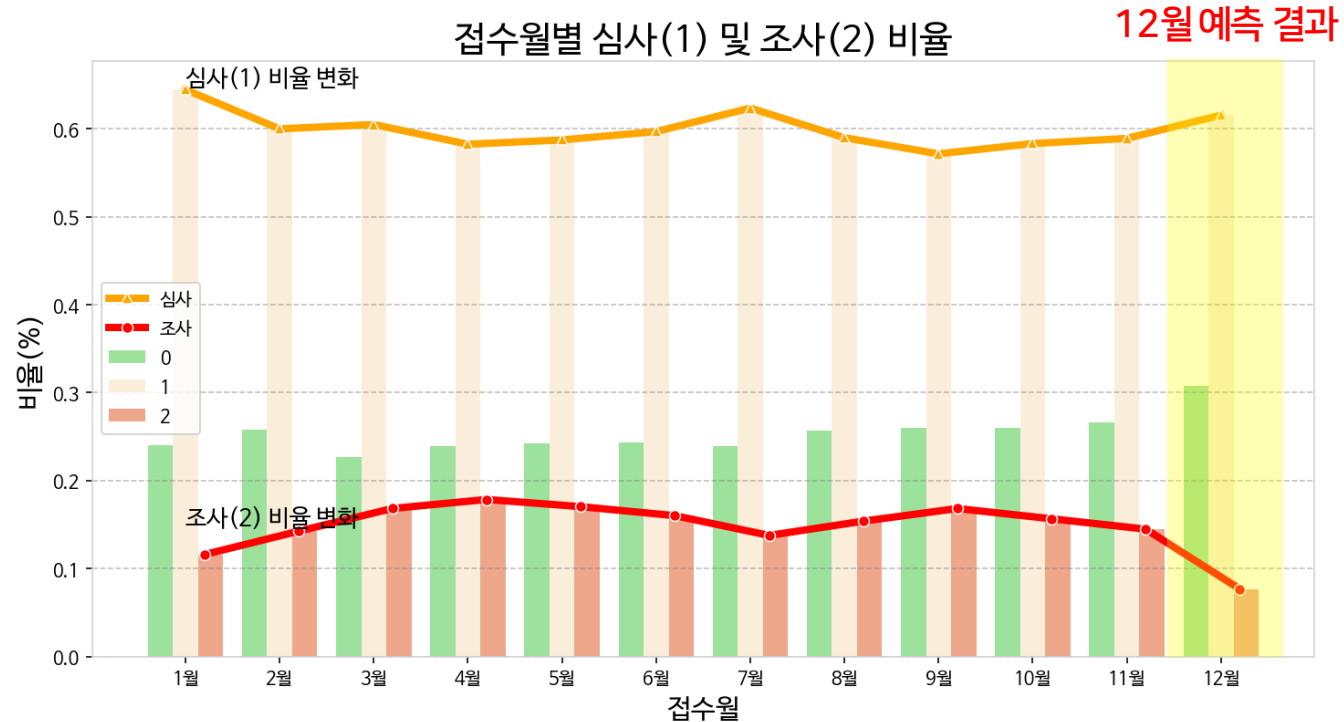
자동지급	226036 (59.8%)
심사	93793 (24.8%)
조사	58099 (15.4%)

자동지급	13584 (61.5%)
심사	6794 (30.8%)
조사	1694 (7.7%)

Public score : 87.074

Model Tuning & Evaluation

모델 평가

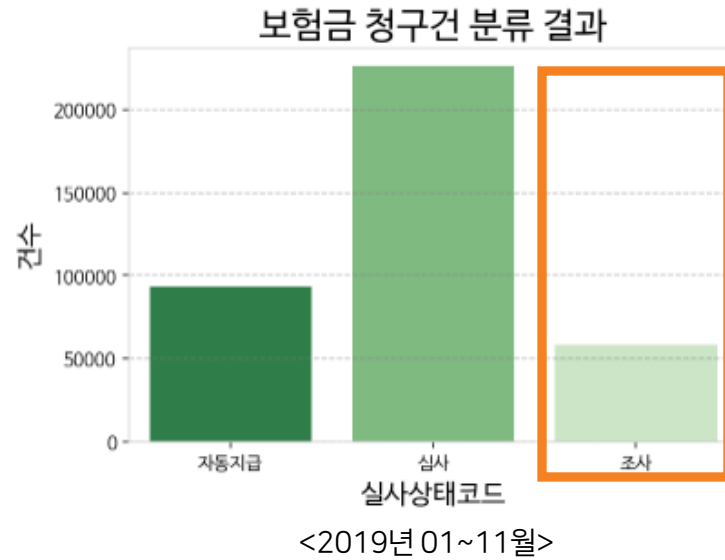


EDA를 통해 파악한 Target 비율 변화 패턴에 따라 12월에는 조사(2)의 비율이 낮아지고 심사(1)의 비율은 높아질 것으로 예상했었다.

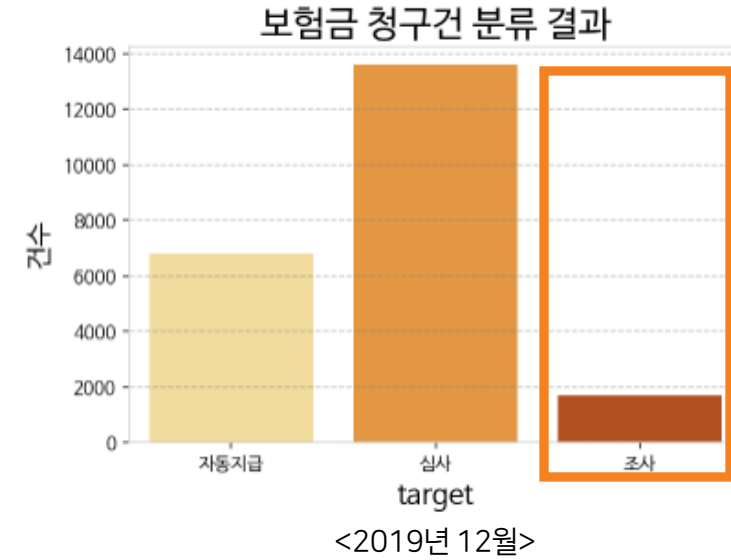
모델링을 통해 12월 보험금 청구 데이터를 예측한 결과가 예측한 패턴에 부합하게 나온 것을 확인할 수 있다.

Model Tuning & Evaluation

모델 평가



예측



주어진 데이터에 비해 예측한 데이터에서 조사의 비율이 낮다.
조사의 데이터가 자동지급이나 심사로 잘못 분류된 것으로 예상된다.

Model Tuning & Evaluation

모델 평가

장점

- 예측률이 일정 확률 이상인 데이터를 학습 데이터에 다시 추가해줌으로서 학습 데이터가 늘어나는 효과를 얻을 수 있었다. f1 score도 유의미하게 상승하였다.
- 모델에서 수정해야 하는 하이퍼 파라미터가 많아 일반적인 gridsearchCV 대신 Bayesian Optimization을 사용함으로써 하이퍼 파라미터를 구하는 시간을 단축시킬 수 있었다.

한계점

- 학습 데이터에 새로운 데이터를 추가해 줄 때 기준이 되는 확률을 임의로 정해야 했다. 즉, 모델링에 주관성이 개입되는 부분이 있다.
- 기존 1~11월 데이터에 비해 조사의 비율을 낮게 예측하는 경향이 있다.

Model Tuning & Evaluation

모델 개선방안

서론

데이터 해석

모델링

결론

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
rf	Random Forest Classifier	0.9787	0.9986	0.9755	0.9789	0.9788	0.9617	0.9618	53.835
et	Extra Trees Classifier	0.9744	0.9971	0.9707	0.9746	0.9744	0.9540	0.9540	72.674
xgboost	Extreme Gradient Boosting	0.9739	0.9983	0.9682	0.9740	0.9739	0.9529	0.9530	337.353
lightgbm	Light Gradient Boosting Machine	0.9734	0.9983	0.9671	0.9734	0.9733	0.9519	0.9520	9.553
catboost	CatBoost Classifier	0.9730	0.9982	0.9669	0.9731	0.9730	0.9512	0.9513	178.825
gbc	Gradient Boosting Classifier	0.9725	0.9981	0.9635	0.9728	0.9724	0.9501	0.9505	395.728
ada	Ada Boost Classifier	0.9669	0.9842	0.9634	0.9669	0.9668	0.9404	0.9404	28.010
dt	Decision Tree Classifier	0.9656	0.9679	0.9639	0.9656	0.9656	0.9382	0.9382	2.025
knn	K Neighbors Classifier	0.9634	0.9872	0.9536	0.9637	0.9633	0.9336	0.9339	9.968
lda	Linear Discriminant Analysis	0.8420	0.9273	0.8111	0.8449	0.8412	0.7154	0.7167	9.370
ridge	Ridge Classifier	0.7763	0.0000	0.6883	0.7794	0.7684	0.5711	0.5790	0.892
nb	Naive Bayes	0.7697	0.9226	0.6525	0.7563	0.7324	0.5647	0.5787	0.530
lr	Logistic Regression	0.6921	0.7861	0.6404	0.6946	0.6927	0.4569	0.4576	6.512
qda	Quadratic Discriminant Analysis	0.5990	0.5000	0.3334	0.3895	0.4488	0.0000	0.0020	5.806

하이퍼파라미터 튜닝이 쉬운 lightgbm을 최종 모델로 사용했으나 pycarat을 이용한 autoML 분석 시 초기 하이퍼파라미터에서 lightgbm 이외에도 점수가 높게 나온 모델이 많았다.

Model Tuning & Evaluation

모델 개선방안

	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
0	0.9821	0.9986	0.9788	0.9822	0.9822	0.9679	0.9679
1	0.9820	0.9986	0.9778	0.9820	0.9820	0.9675	0.9675
2	0.9798	0.9986	0.9765	0.9799	0.9799	0.9637	0.9637
3	0.9797	0.9986	0.9765	0.9798	0.9797	0.9634	0.9634
4	0.9823	0.9988	0.9780	0.9823	0.9823	0.9680	0.9681
5	0.9805	0.9986	0.9767	0.9806	0.9805	0.9649	0.9649
6	0.9815	0.9987	0.9791	0.9816	0.9816	0.9668	0.9668
7	0.9807	0.9985	0.9776	0.9808	0.9807	0.9653	0.9653
8	0.9797	0.9984	0.9764	0.9798	0.9797	0.9635	0.9635
9	0.9810	0.9985	0.9776	0.9810	0.9810	0.9658	0.9658
Mean	0.9809	0.9986	0.9775	0.9810	0.9809	0.9657	0.9657
SD	0.0010	0.0001	0.0009	0.0009	0.0010	0.0017	0.0017

특히 Extra Tree Classifier, Random Forest Classifier, 및 lightgbm Classifier을 soft vote ensemble했을 때 F1 score가 단일 모델보다 0.4~0.8정도 향상되는 결과를 볼 수 있었다.

따라서 타 모델과 앙상블 시도 시 더 높은 정확도를 얻을 수 있을 것으로 예상된다.

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
0	Voting Classifier	0.9815	0.9987	0.9783	0.9816	0.9815	0.9668	0.9668

PART

4

서론

데이터 해석

모델링

결론

비즈니스 활용 방안

보험금 청구 결과 예측 서비스

보험금 청구 결과 예측 서비스

서비스 개요

고객이 보험금 청구를 접수하기 전에 머신러닝 모델을 활용해 고객의 청구 데이터를 분석하여 예측된 청구 분류 결과를 알려주는 서비스.

기대효과

- 조사로 분류될 시 향후 현장조사가 발생할 수 있다는 사실을 미리 공지한다. 또한 이런 상황이 자주 일어날 경우 보험사기 위험고객으로 분류되어 이후에 보험금 지급이 어려워질 수 있음을 경고한다. 이를 통해 고객이 보험금 청구를 신중하게 할 수 있도록 하고, 현장조사에 필요한 인력 낭비를 방지할 수 있다.
- 머신러닝을 통해 자동지급건을 자동 분류하고 보험금을 자동 지급함으로써 불필요한 서류 분석 과정을 단축할 수 있다.

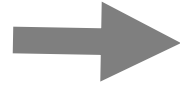
비즈니스 활용 방안

보험금 청구 결과 예측 서비스

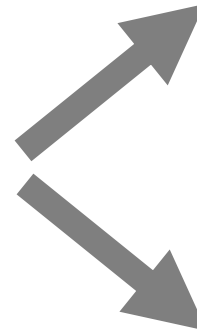
예측 서비스 프로세스



개인정보 및
가입상품정보 확인



제출문서 파악 후
model을 통한
청구 분류 예측
(문서 파악을 위한
스캔 이미지 분석
알고리즘 필요)



자동지급 분류시
자동 보험 청구 접수 및
보험금 지급



조사 분류시 추가 필요
서류 및 현장조사가
필요할 수도 있음을 알림.

비즈니스 활용 방안

보험금 청구 결과 예측 서비스(활용 예상도)

1 보험자 개인정보 및 계약정보 확인

The screenshot shows the '보험 청구 예측' (Insurance Claim Prediction) page. Under the heading '1 피보험자 및 계약정보 확인' (1. Policyholder and Contract Information Confirmation), there are three input fields: '이름을 입력해주세요.' (Please enter your name.), '생년월일을 입력해주세요.' (Please enter your date of birth.), and '고객번호를 입력해주세요.' (Please enter your customer number.). An orange '확인' (Confirm) button is at the bottom.

2 보험금 청구 서류 등록

The screenshot shows the '보험 청구 예측' (Insurance Claim Prediction) page. Under the heading '2 청구사항 입력' (2. Claim Information Input), there is a dropdown menu for '가입 보험' (Enrolled Insurance), a blue button for '보험금 청구양식 다운로드' (Download Insurance Claim Form), and an input field for '파일을 첨부하세요.' (Attach file.). An orange '서류 업로드' (Upload Document) button is at the bottom.

3-1 청구 결과 확인 #1 : 자동지급

The screenshot shows the '보험 청구 예측' (Insurance Claim Prediction) page. Under the heading '3 청구 결과 확인' (3. Claim Result Confirmation), the text says '자동지급으로 분류됩니다.' (Categorized as automatic payment.). Below this, it explains: '자동지급으로 분류될 경우 빠른 시간 내에 보험금 지급이 이루어지며, 추가적인 조사는 없습니다.' (When categorized as automatic payment, the insurance claim will be paid within a short time, and no further investigation is required.). It then asks '합수하시겠습니까?' (Do you agree?). At the bottom are two orange buttons: '보험금 청구하기' (Apply for Insurance Claim) and '돌아가기' (Go Back).

3-2 청구 결과 확인 #2 : 조사

The screenshot shows the '보험 청구 예측' (Insurance Claim Prediction) page. Under the heading '3 청구 결과 확인' (3. Claim Result Confirmation), there is a warning icon and the text '조사로 분류됩니다.' (Categorized as investigation.). Below this, it explains: '조사로 분류될 경우 현장조사가 발생 할 수 있으며, 번민하게 사례가 발생하면 보험사가 위험 고객으로 분류될 수 있습니다.' (When categorized as investigation, a field investigation may occur, and if a case occurs, the insurer may classify you as a high-risk customer.). It then asks '그래도 접수하시겠습니까?' (Do you still want to file a claim?). At the bottom are two orange buttons: '보험금 청구하기' (Apply for Insurance Claim) and '돌아가기' (Go Back).

감사합니다

Thank You!

