

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



성명 : 서정민
학교 : 송실대학교

학년 : 3학년
학과 : 산업정보시스템공학과



성명 : 이혜원
학교 : 고려대학교

학년 : 석사과정
학과 : 통계학과

CONTENTS

01 서론

- 주제 이해
- 문제 상황 및 배경 파악

02 데이터 해석

- 데이터 이해
- EDA

03 모델링

- Feature Engineering
- Model tuning & Evaluation

04 비즈니스 활용방안

- 분류 정확도 향상
- 개인화 및 고객 서비스

분석 과정

서론

주제 이해 및
분석 목표 설명

데이터 해석

모델 정확도 및
분석 목표
달성을 위한
데이터 분석

모델링

EDA를 활용해
최적의 모델
도출

활용방안

데이터 분석내용 및
모델을 활용하여
비즈니스
활용방안 제시

PART

1

서론

데이터 해석

모델링

비즈니스 활용방안

주제 이해

보험금 청구 프로세스

서론

데이터 이해

모델링

비즈니스 활용방안

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 이해

모델링

비즈니스 활용방안

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.

문제상황 및 배경 파악

보험 사기의 증가

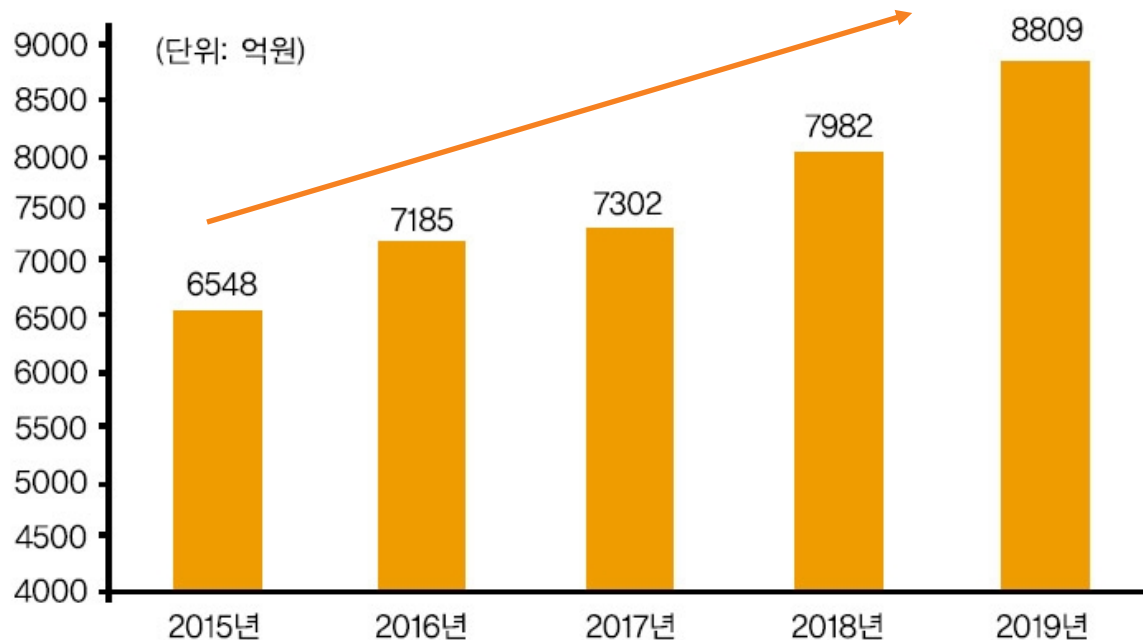
서론

데이터 이해

모델링

비즈니스 활용방안

〈 보험사기 적발금액 추이 〉



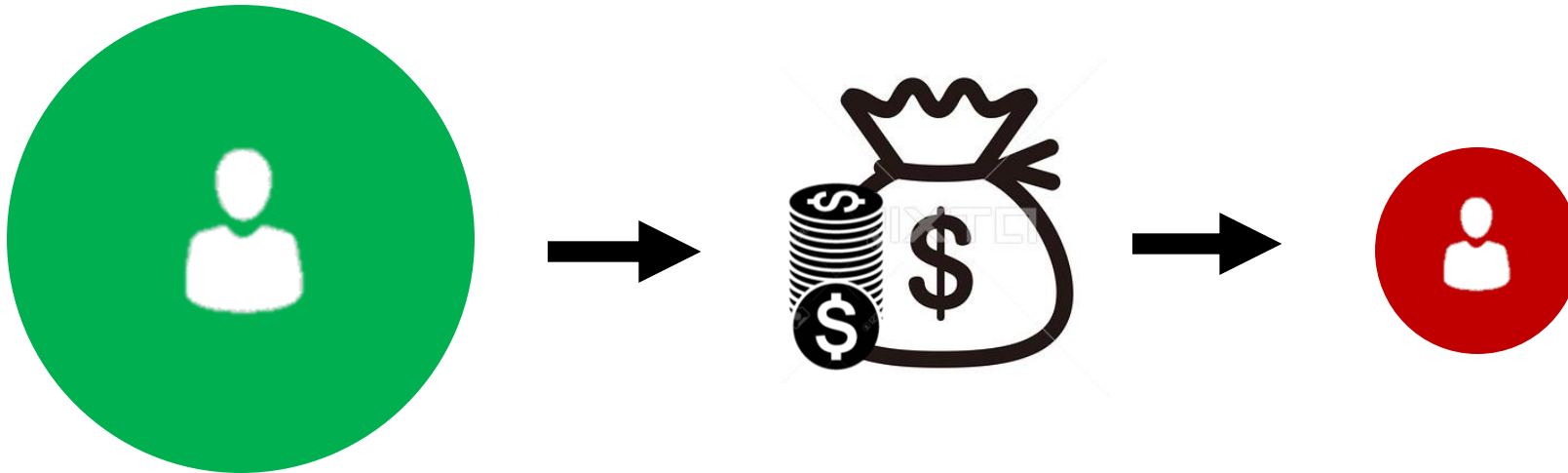
* 자료: 금융감독원

- 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

분석 방향 및 목표

보험은 다수의 고객으로부터 보험료를 모아 소수의 고객에게 보험금을 지급하는 구조.
따라서 보험사기 등으로 이를 악용하는 고객들로 인해 규정을 잘 지키고 있는
다수의 선량한 고객들이 피해를 받는 문제가 존재한다.



문제상황 및 배경 파악

분석 방향 및 목표

서론

데이터 이해

모델링

비즈니스 활용방안

소수 고객들의 모럴 해저드 때문에 기업의 손실이 발생하고,
이는 고스란히 선량한 고객들의 피해로 이어진다.



기업



고객

- 악의의 고객을 거르기 위한 인력 및 비용 발생
- 거르지 못한 악의의 고객에게 지급되는 과도한 보험금

보험료가 상승하면서
다수의 선량한 고객들 또한 피해

문제상황 및 배경 파악

분석 방향 및 목표

따라서 이 문제를 해결하는 것은 단기적으로는 회사의 손실을 줄이는 데 도움이 되지만 장기적으로는 고객의 이익으로 돌아갈 것이라 보았다.



그러므로 데이터 분석 시 단순히 모델의 정확도를 높이는 것과 동시에 **어떻게 하면 악의의 고객을 잘 분류하면서 선의의 고객에게 이익이 갈 수 있는지에 주안점을 두었다.**

PART 2

서론

데이터 해석

모델링

비즈니스 활용방안

데이터 이해

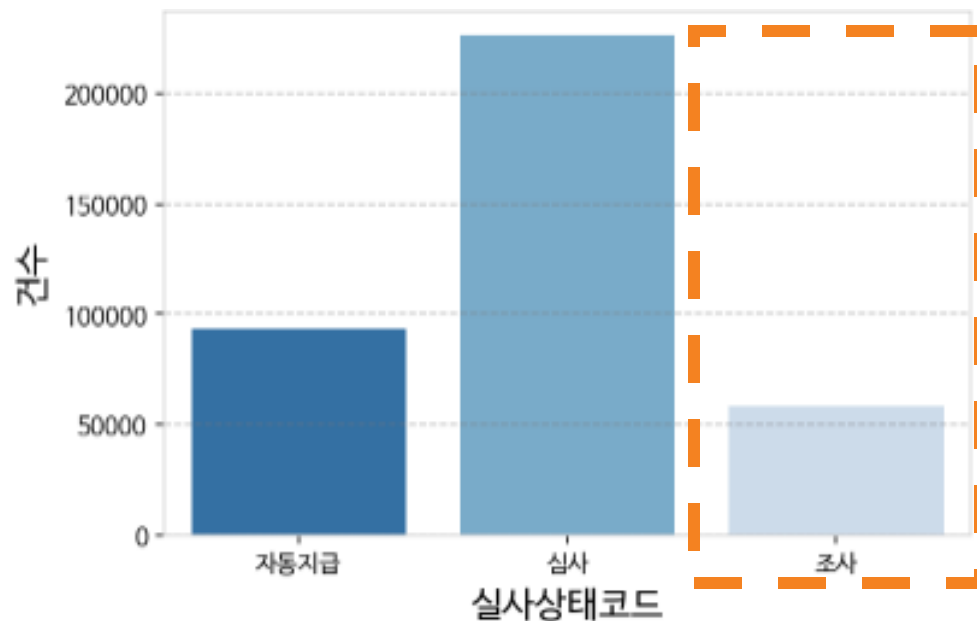
데이터 변수 탐색

서론

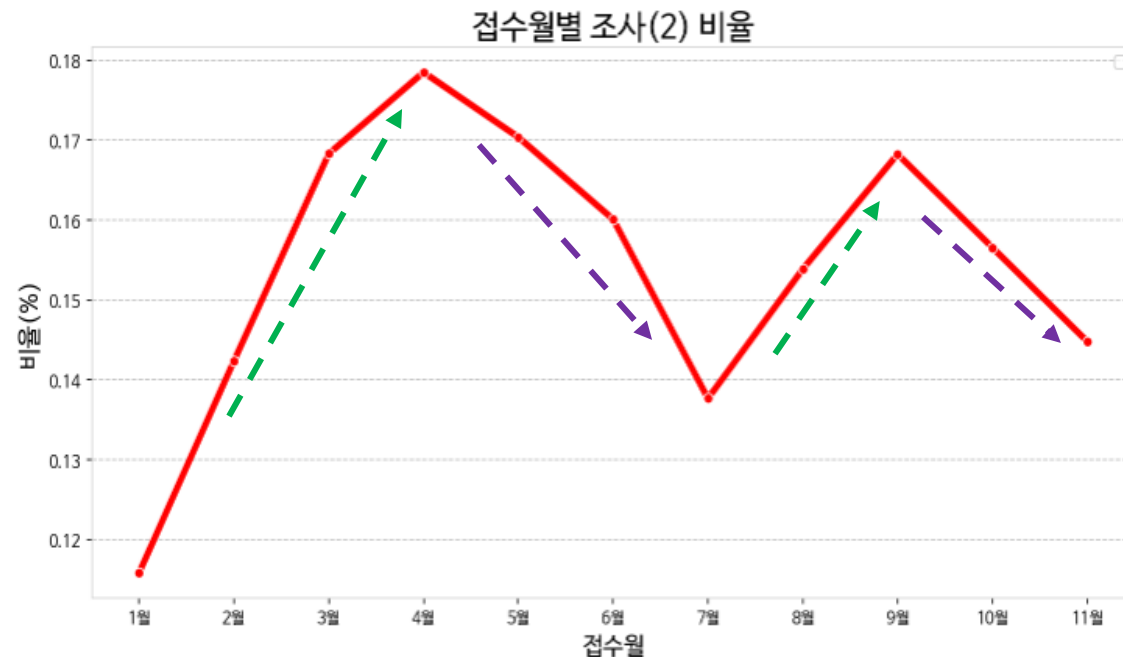
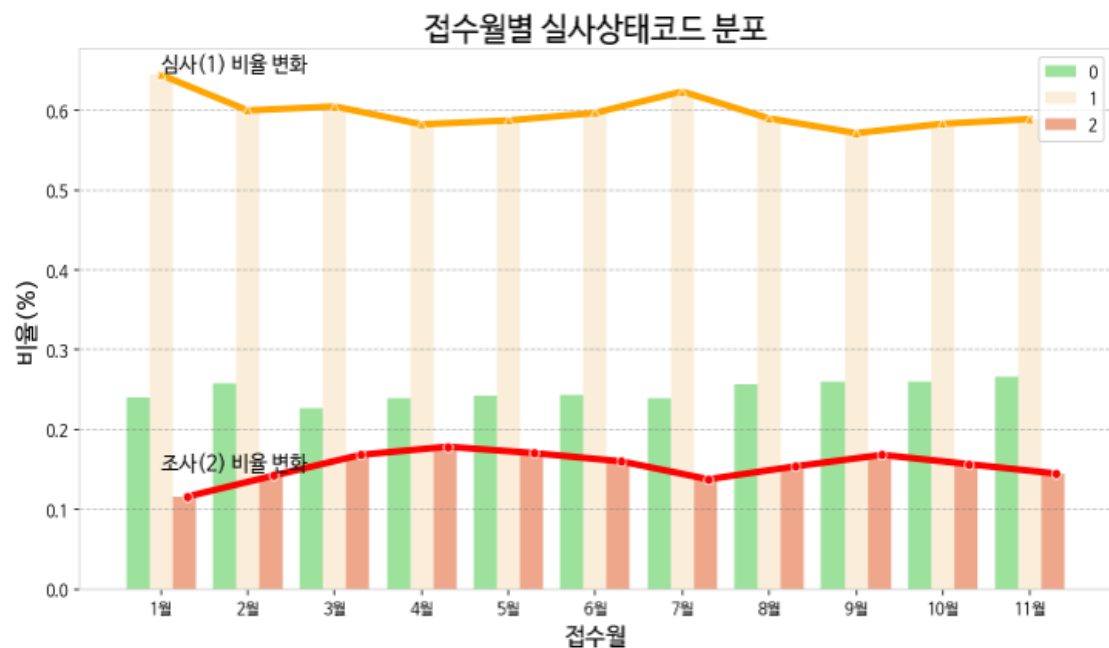
데이터 해석

모델링

비즈니스 활용방안

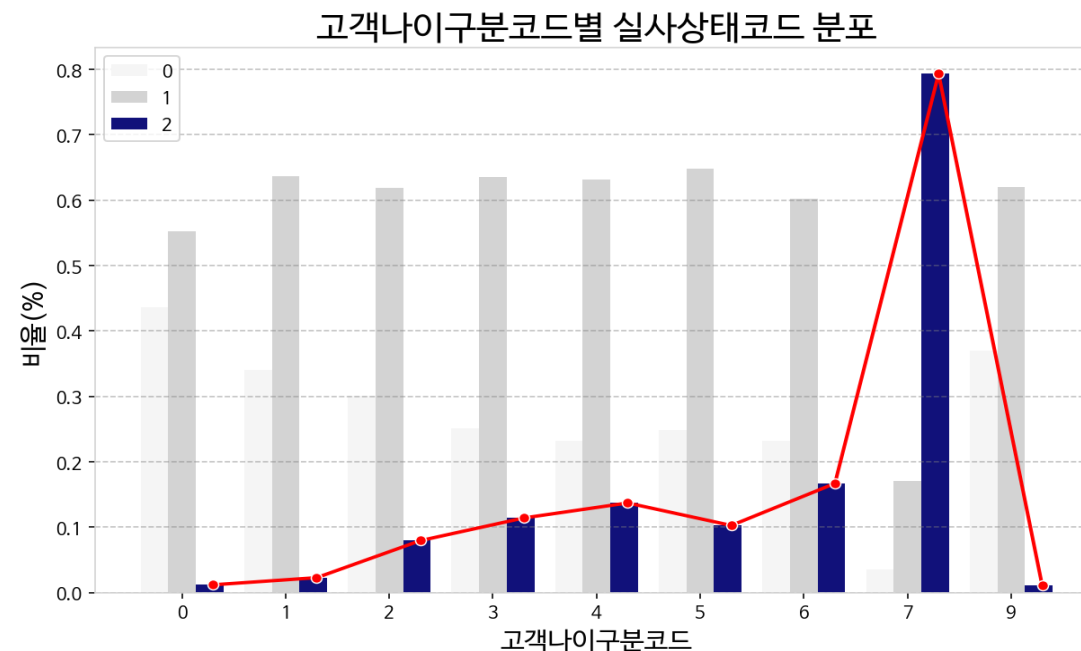
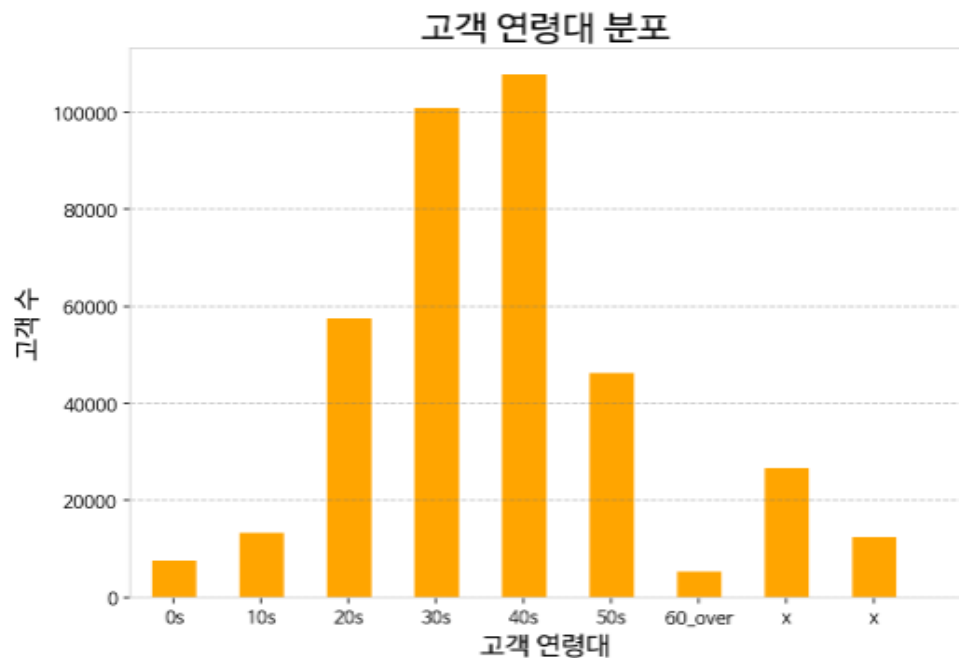


회사와 고객 모두에게 이익이 돌아가기 위해서는 보험사기 등의 가능성이 높은 '조사' 데이터를 잘 분류해야 한다.
따라서 어떤 특성을 가질 때 '조사' 카테고리 분류되고, 어떻게 하면 이러한 고객들을 '조사'로 정확하게 예측할 수 있을가에 주안점을 두고 데이터 해석 및 EDA를 진행하였다.



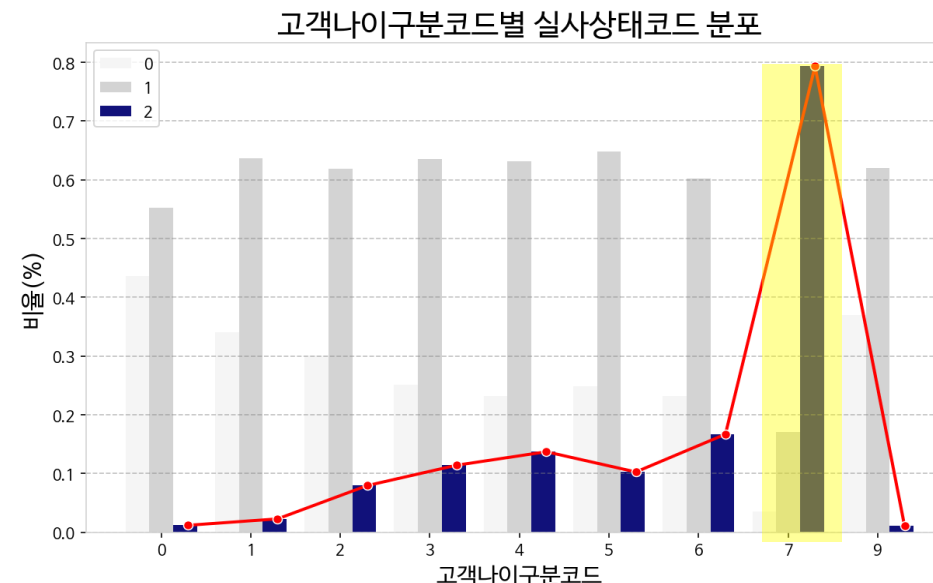
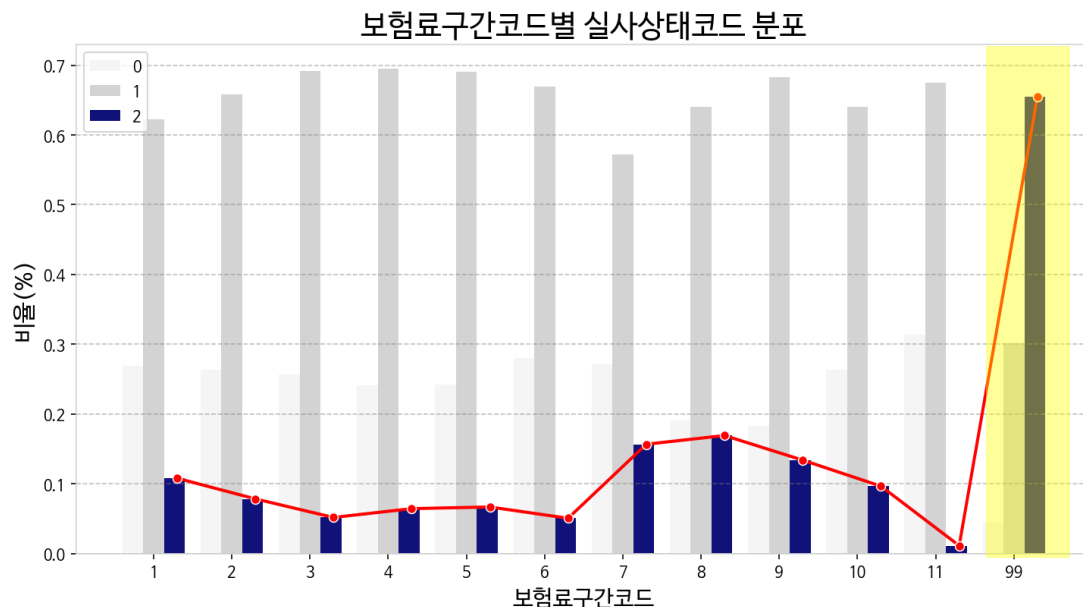
접수월에 따라 실사상태코드의 분포가 달라진다. 심사(1)와 조사(2)는 일년동안 상승과 하락을 반복했고, 특히 자동지급의 비율은 매달 일정한 반면 심사와 조사의 비율은 서로 상반되는 추세를 보였다.

따라서 **보험금 청구 분류 결과에는 계절성이 중요하게 반영된다고 판단했다.**



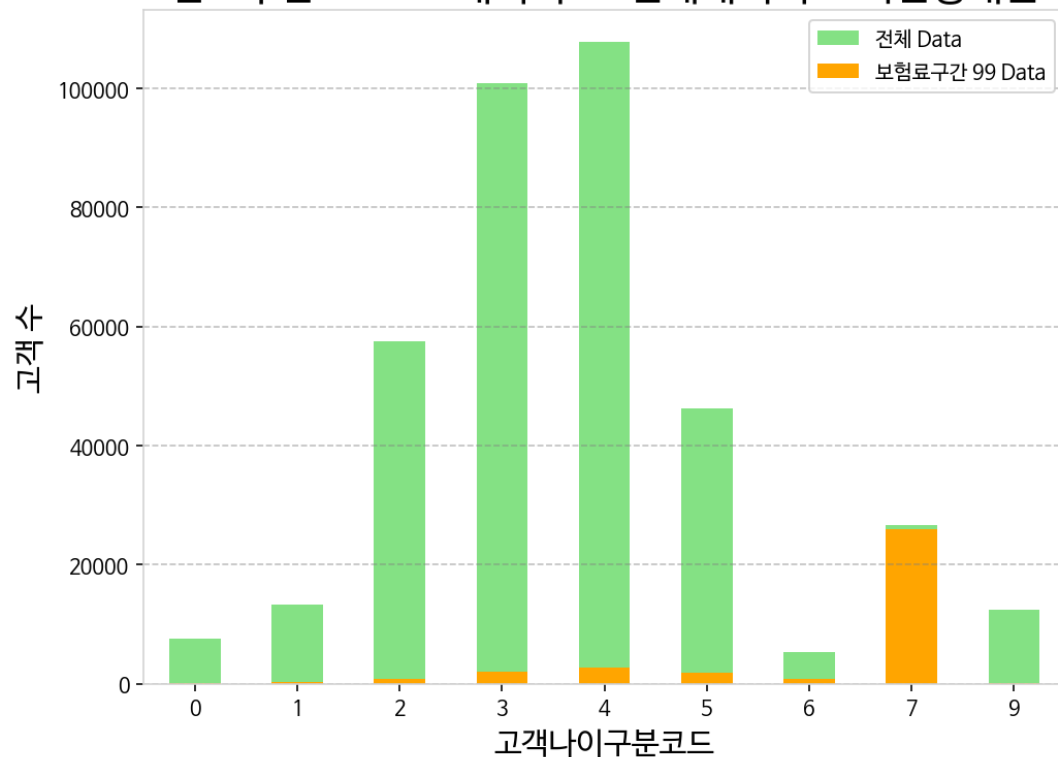
고객 연령대는 20~50대가 대다수이며 60대 이상이 가장 작은 비율을 보였다. 대체적으로 연령이 높아질수록 조사(2)의 비율이 높아지나, 특이하게도 고객나이구분코드가 7일 때 조사 비율이 다른 고객나이구분코드에 비해 압도적으로 높은 것을 볼 수 있다.

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



보험료구간코드가 99(unknown)값을 가질 때 조사의 비율이 압도적으로 높은 것을 확인할 수 있었다. 고객나이구분코드가 7일 때도 조사의 비율이 매우 높았는데 이를 통해 고객나이구분코드가 7인 데이터와 보험료구간코드가 99인 데이터가 관련이 있다고 판단하였다.

보험료구간코드 99 데이터 vs 전체데이터 고객연령대분포

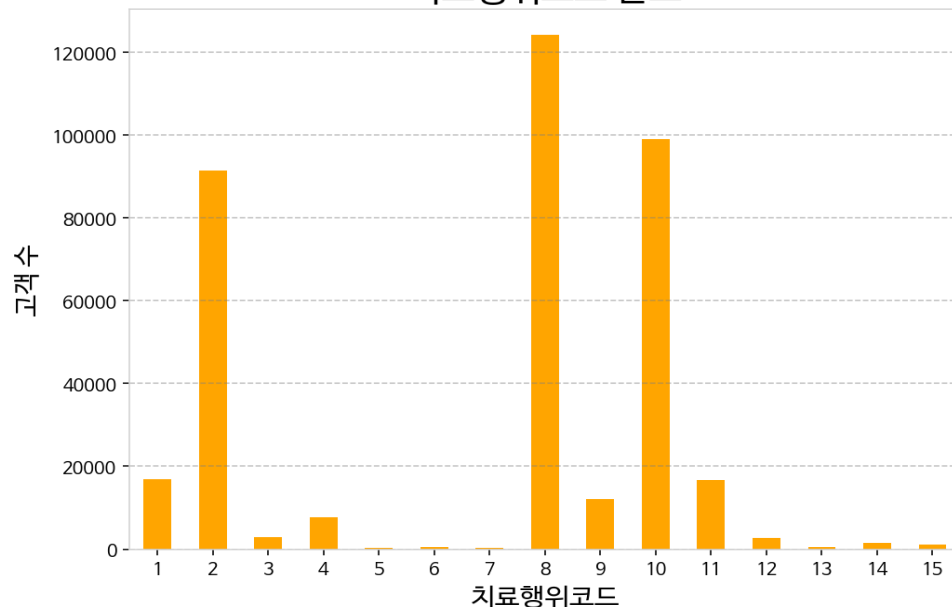


보험료구간코드가 99인 데이터는 대부분 고객나이구분코드 7에 속했다.

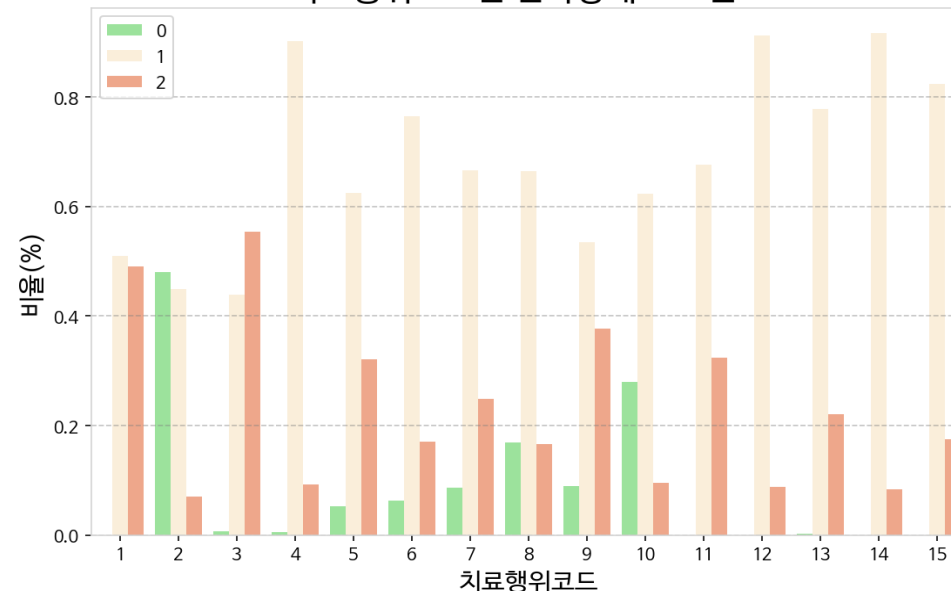
마찬가지로 고객나이구분코드 7에 속하는 고객의 99%는 보험료구간코드 99에 해당되었다.

즉, **고객나이구분코드, 보험료구간코드, 청구금액구간코드 및 청구일계약일간구분코드 중 하나라도 unknown 데이터라면 다른 변수들도 대부분 unknown값을 가지게 된다.**

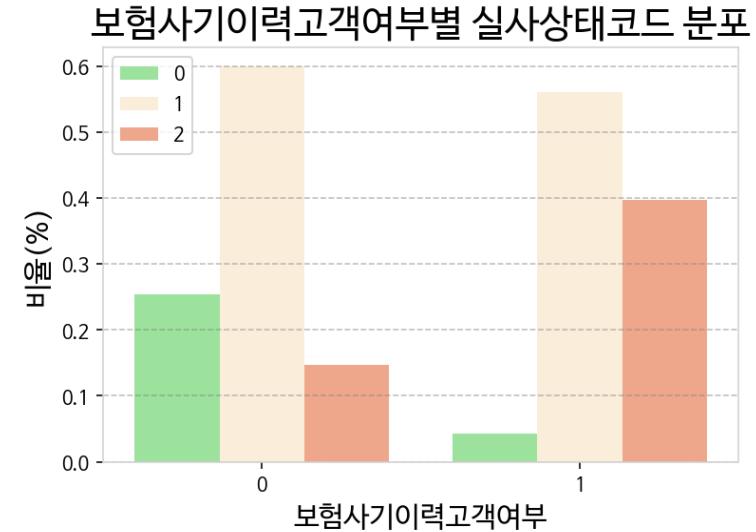
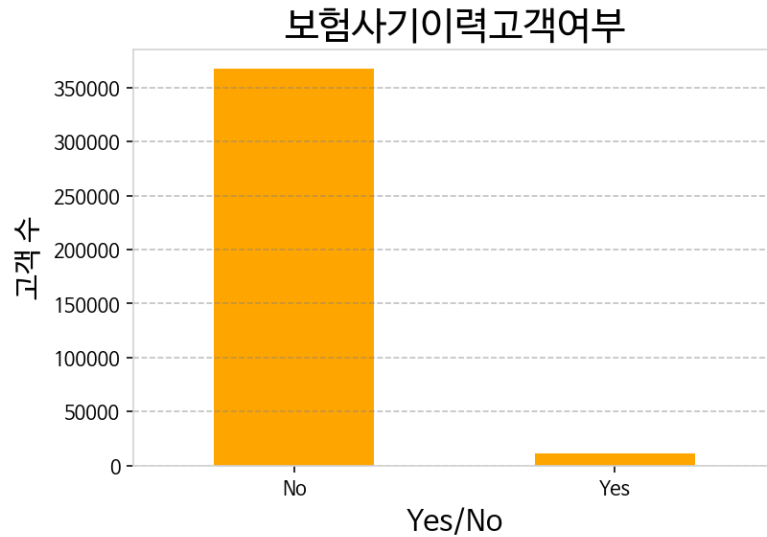
치료행위코드 분포



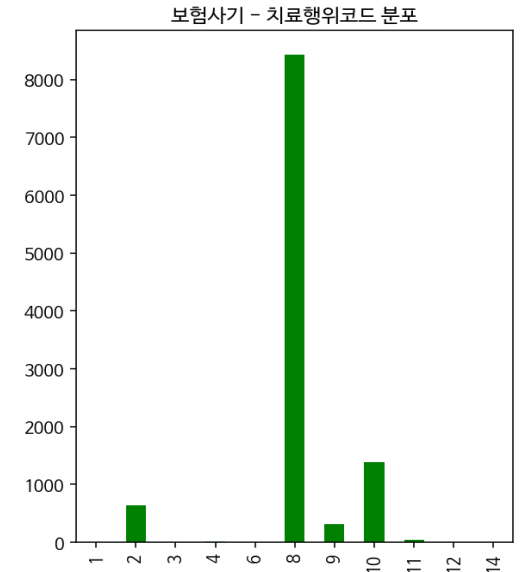
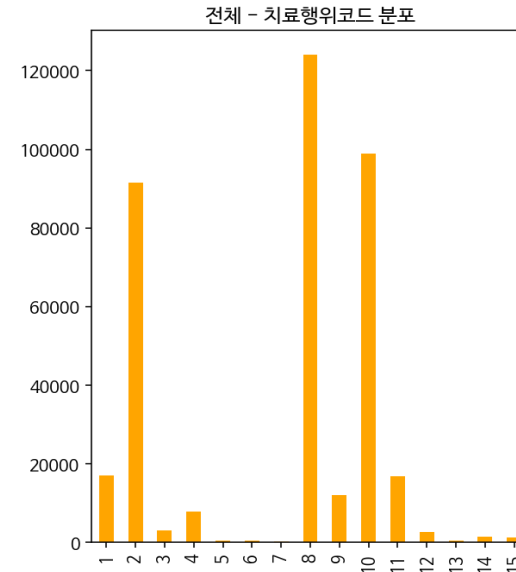
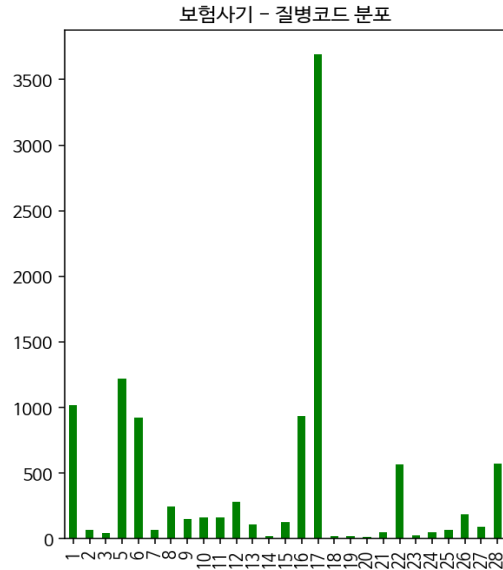
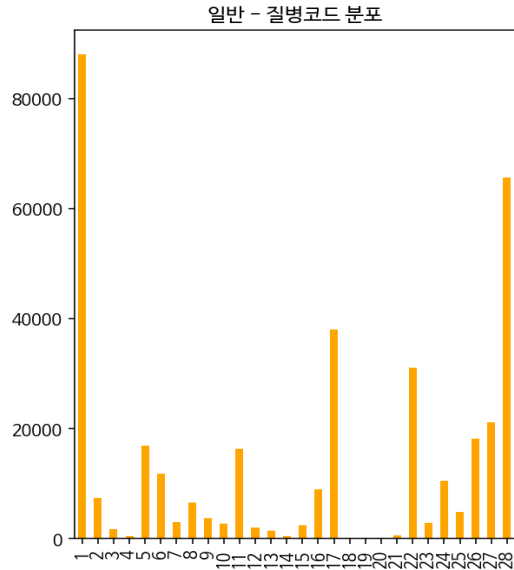
치료행위코드별 실사상태코드 분포



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.



보험사기이력이 있는 고객의 비율은 매우 낮다. 그러나 보험사기 이력이 있을 경우 그렇지 않을 때와 비교해 조사의 비율이 매우 높았다.
따라서 **보험사기이력이 실사상태코드에 영향을 미친다**고 예상할 수 있다.



보험사기이력이 있는 고객의 경우 청구 질병 및 치료행위 등에서 일반 고객과는 다른 분포를 보였다.
ex) 보험사기 이력이 있을 때 관절염 청구 비율이 매우 높았으며, 입원 치료만 받은 비율은 87.4%에 달했음.

- **접수년월**별로 실사상태코드의 분포가 변화하는 양상을 띄었기 때문에 계절적 요소를 추가해야 할 것으로 판단되었다.
- **고객나이구분코드, 보험료구간코드, 가입금액구간코드 및 청구일계약일간구분코드**의 unknown 값은 서로 관련성이 높았다. 따라서 이러한 unknown여부를 나타내는 새로운 변수가 필요한 것으로 판단되었다.
- **질병치료행위**별로 실사상태코드 분포가 다르게 나타났다. 따라서 질병치료행위(입원, 통원, 수술, 진단) 여부를 나타내는 새로운 변수를 만들어 보고자 하였다.
- **보험사기이력**이 있을 경우 조사로 분류될 확률이 높았다. 따라서 차후 모델 활용방안을 도출할 때 보험사기이력이 있는 고객의 특성을 파악해 반영하는 것이 중요하다고 생각되었다.

PART 3

서론

데이터 해석

모델링

비즈니스 활용방안

Feature Engineering

변수 선택

계절성



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

Feature Engineering

변수 선택

서론

데이터 해석

모델링

비즈니스 활용방안

의료기관구분코드



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

Feature Engineering

변수 선택

서론

데이터 해석

모델링

비즈니스 활용방안

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

시도하였으나 성능 저하로 수정하지 않은 변수

Unknown
값 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

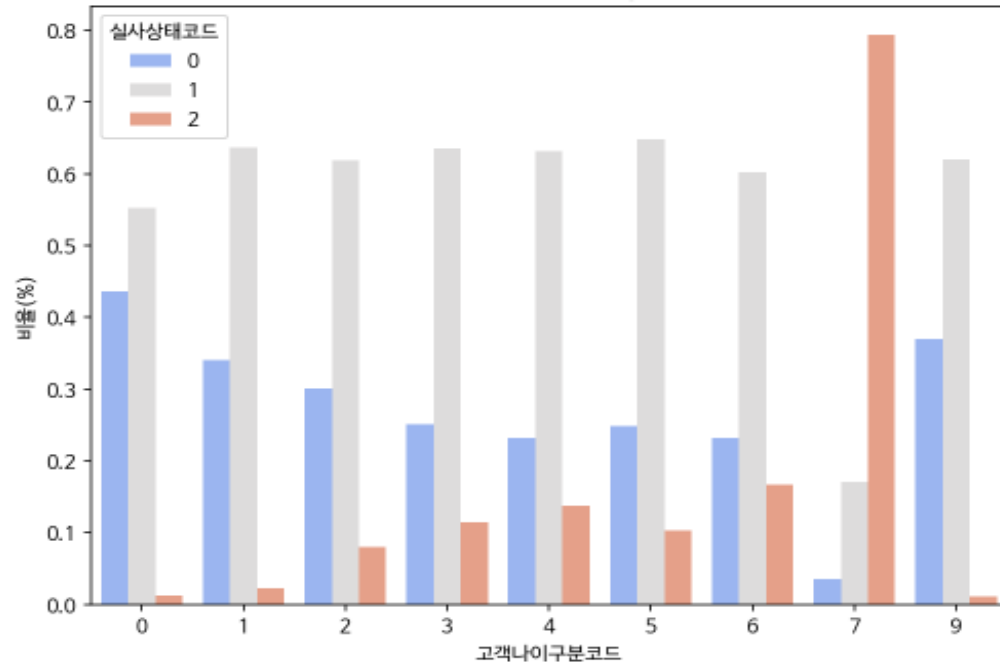
고객나이
구분코드

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

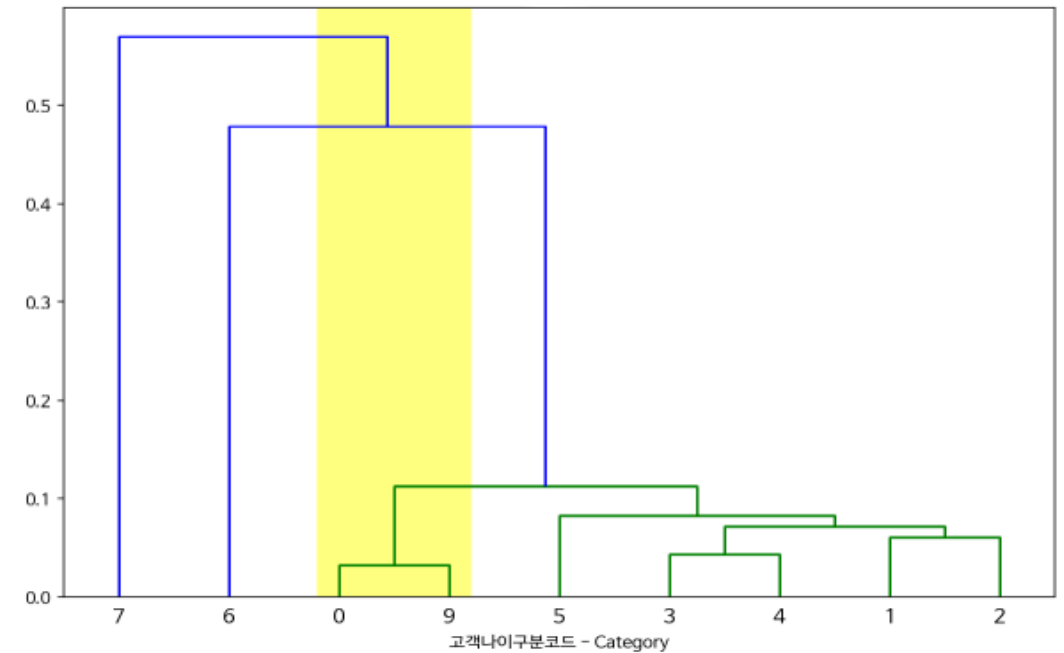
Feature Engineering

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

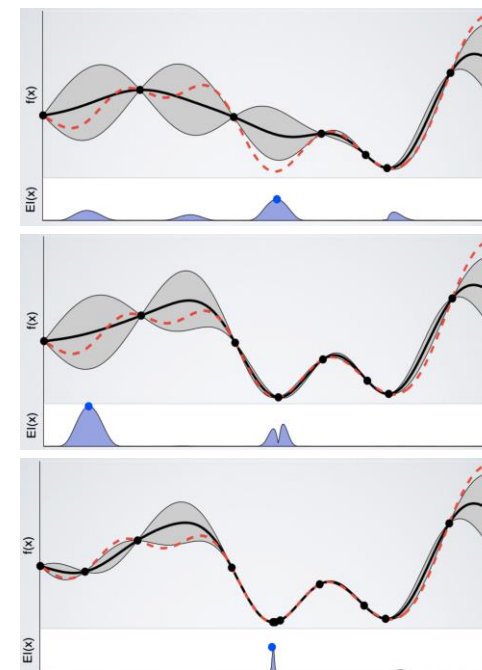
데이터 해석

모델링

비즈니스 활용방안

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



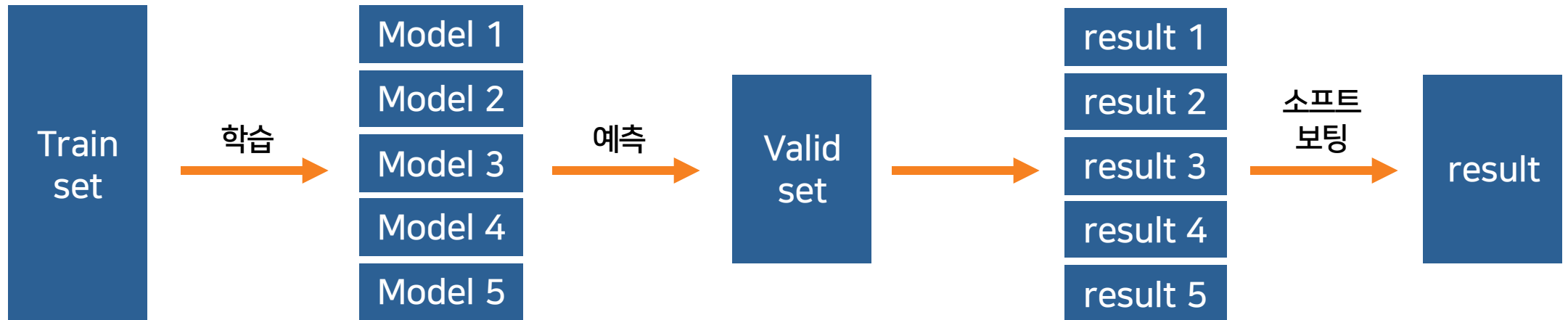
하이퍼 파라미터 튜닝을 위해 Bayesian Optimization을 사용했다.

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

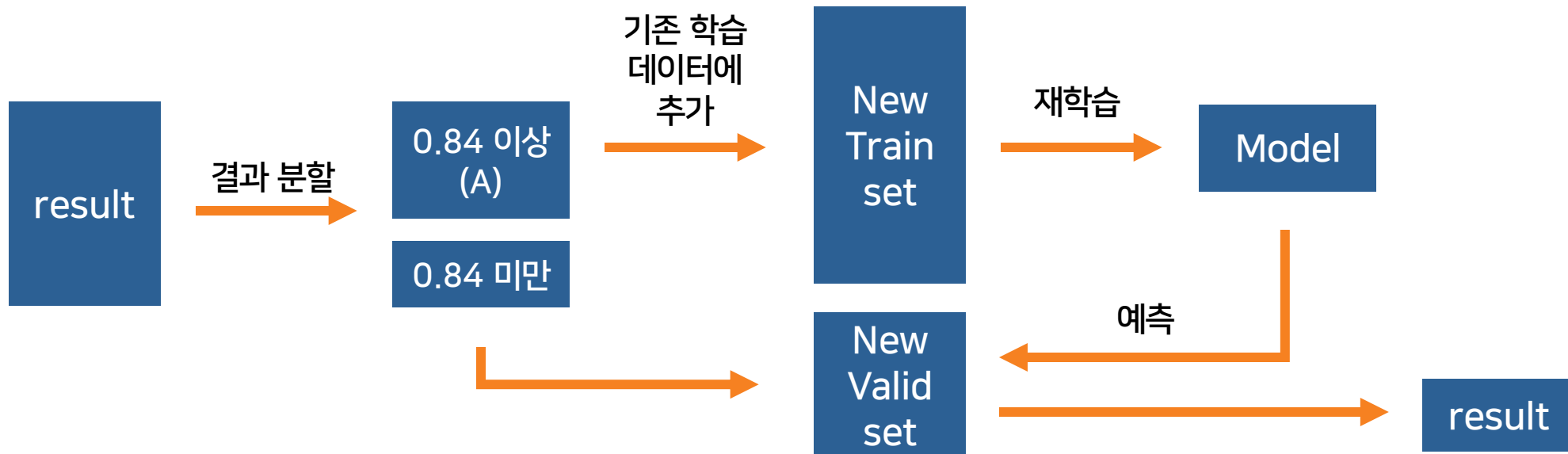


5개의 결과를 소프트 보팅함으로써 약 0.3~0.5점의 f1 score 상승을 얻을 수 있었다.

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

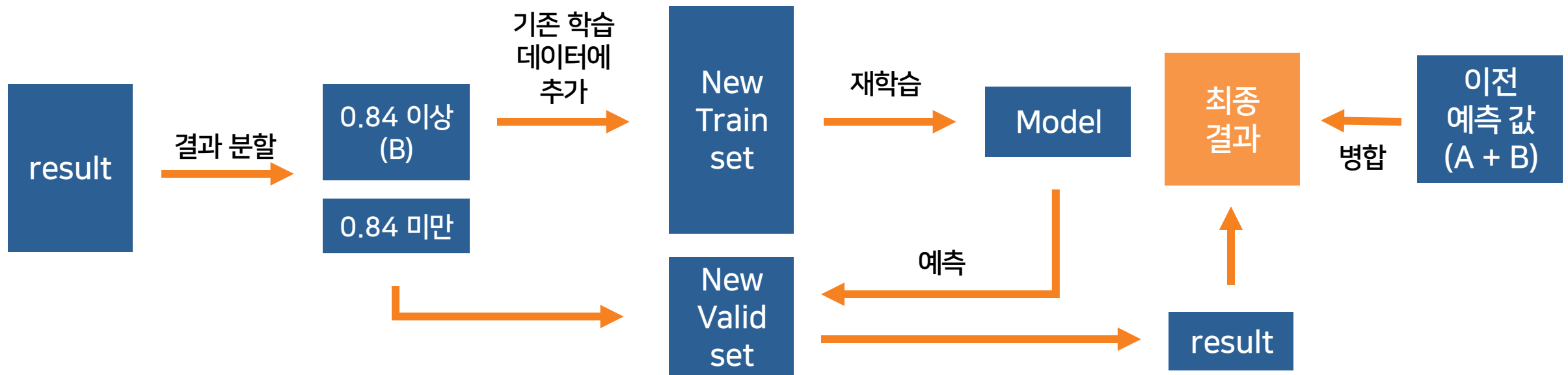


최대확률이 0.84 이상인 데이터를 학습 데이터에 추가해주면서 학습 데이터가 증가하는 효과가 나타나 예측 정확도가 상승했다.

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

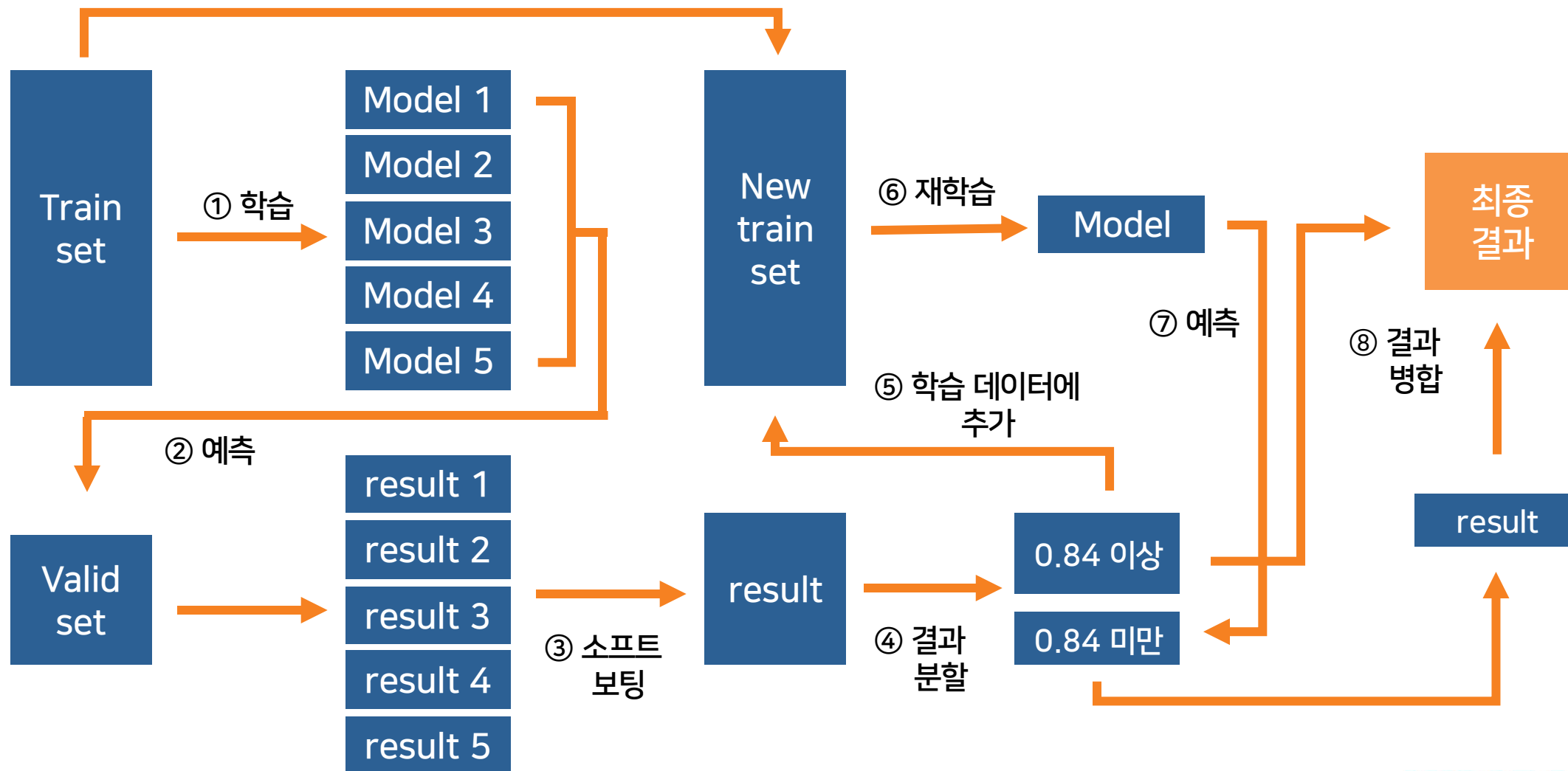
모델 구조화

서론

데이터 해석

모델링

비즈니스 활용방안



Model Tuning & Evaluation

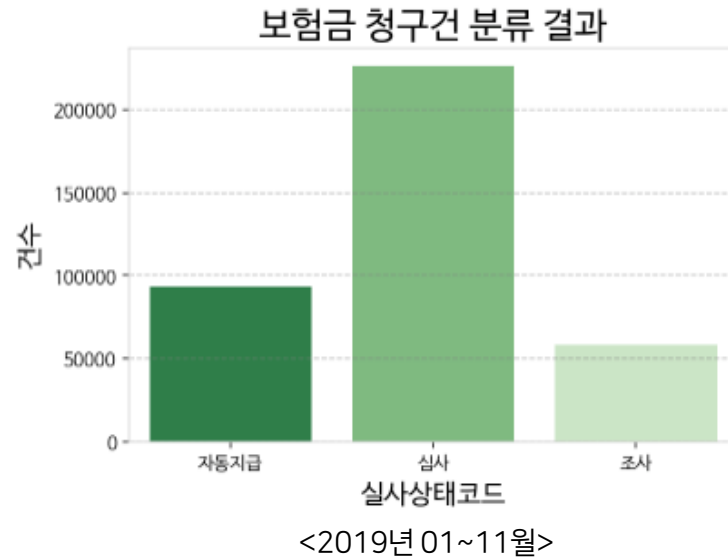
모델 평가

서론

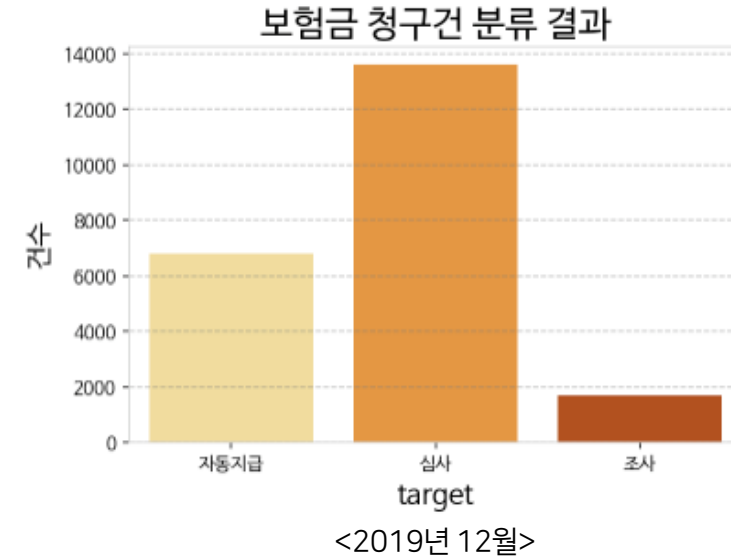
데이터 해석

모델링

비즈니스 활용방안



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



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

Public score : 87.074

Model Tuning & Evaluation

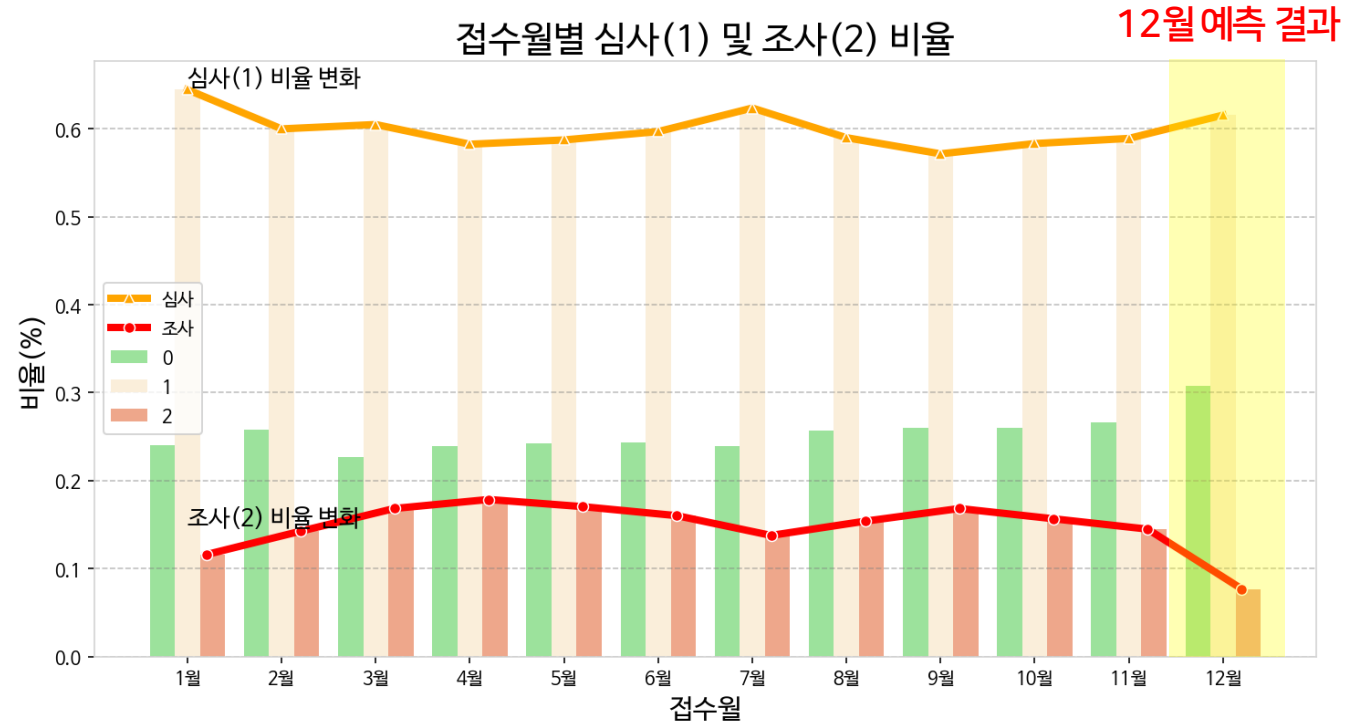
모델 평가

서론

데이터 해석

모델링

비즈니스 활용방안



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.

PART

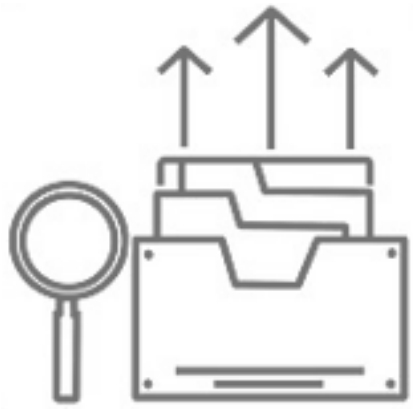
4

서론

데이터 해석

모델링

비즈니스 활용방안



1. 보험금 청구 분류 모델의 **정확도(특히 조사)를 높여** 악의의 고객을 잘 분류해 내는 것



2. 모델 및 데이터 특성을 파악하여 **선의의 고객에게 혜택**이 돌아가는 것

두개의 큰 틀에서 우리가 앞서 생각했던 문제점들을 해결할 수 있다고 판단하였다.

분류 정확도 향상

1. 데이터 재 라벨링

데이터 분석 과정에서 학습 데이터 내에서 평가한 f1 score와 12월 데이터를 예측했을 때의 f1 score간의 차이가 큰 것을 알 수 있었다.

학습 데이터 내에서
측정한 f1 score

97.503

10.86점 차이



실제 데이터로
측정한 f1 score

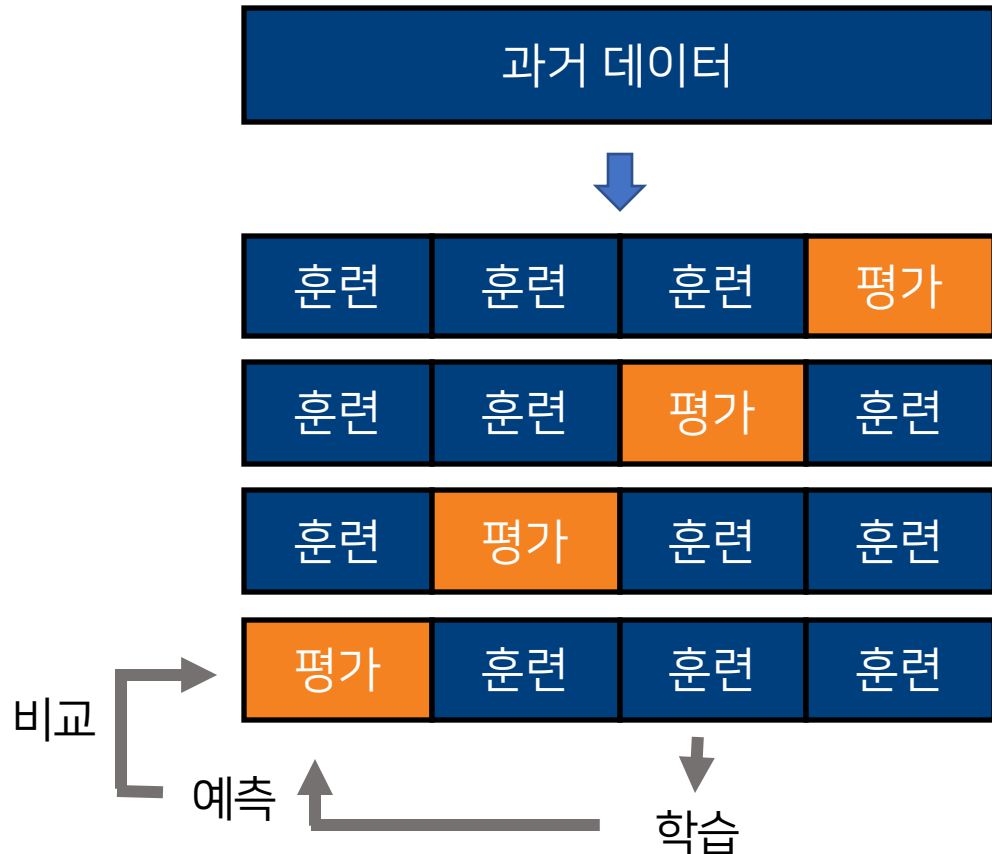
86.64

오버피팅의 문제뿐만 아니라
라벨링의 문제(과거 데이터의 부정확성)는 아니었을까?

분류 정확도 향상

1. 데이터 재 라벨링

지금까지 보험청구 분류는 사람이 직접 수행했기 때문에 오류 데이터가 존재할 수 있다.
 (ex. 특정 청구 건을 자동 지급으로 분류했었으나 실제로는 조사가 필요했던 경우)
 이러한 과거 분류 건의 오류를 찾아내는 것 또한 모델 정확도를 높이는데 중요하다고 판단했다.

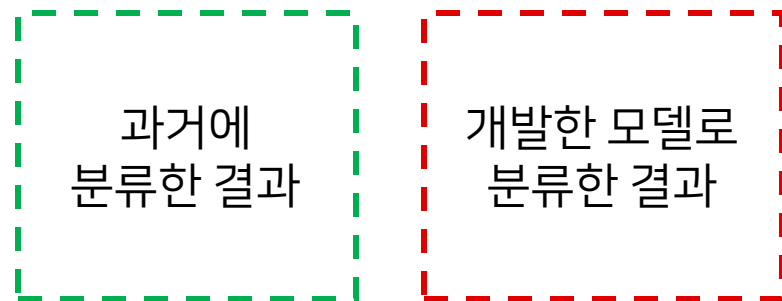


훈련 데이터로 모델을 학습한 후, 평가 데이터의
 모델 예측 결과를 실제 target 값과 비교
 예측 결과와 실제 target 값 간에 차이가 있을
 경우 과거 분류과정에 오류가 있었을 가능성이
 있다고 판단하고 재검토 수행

분류 정확도 향상

1. 데이터 재 라벨링

AB 테스트



과거 오분류된 데이터 확인

활용방향 및 기대효과

1. 모델의 개선 방향 설정 가능

- 오분류된 데이터의 특징을 더 잘 잡아낼 수 있는 모델 생성
- 과거 데이터의 target 값 수정으로 모델의 정확도 향상

2. 잘못 분류된 데이터의 특징 추출 및 분석 가능

- 악의의 고객이 분류 프로세스를 교묘하게 통과하여 조사 건이 자동지급으로 분류된 사례가 없는지 파악하고 보험사기 예방에 이용

분류 정확도 향상

2. 이분류 모델링

자동지급/심사/조사 중 조사를 제대로 분류하는 게 가장 중요하다고 생각되어
기존 자동지급/심사/조사의 3분류가 아닌 자동지급+심사/조사의 2분류 모델을 생성하였다.



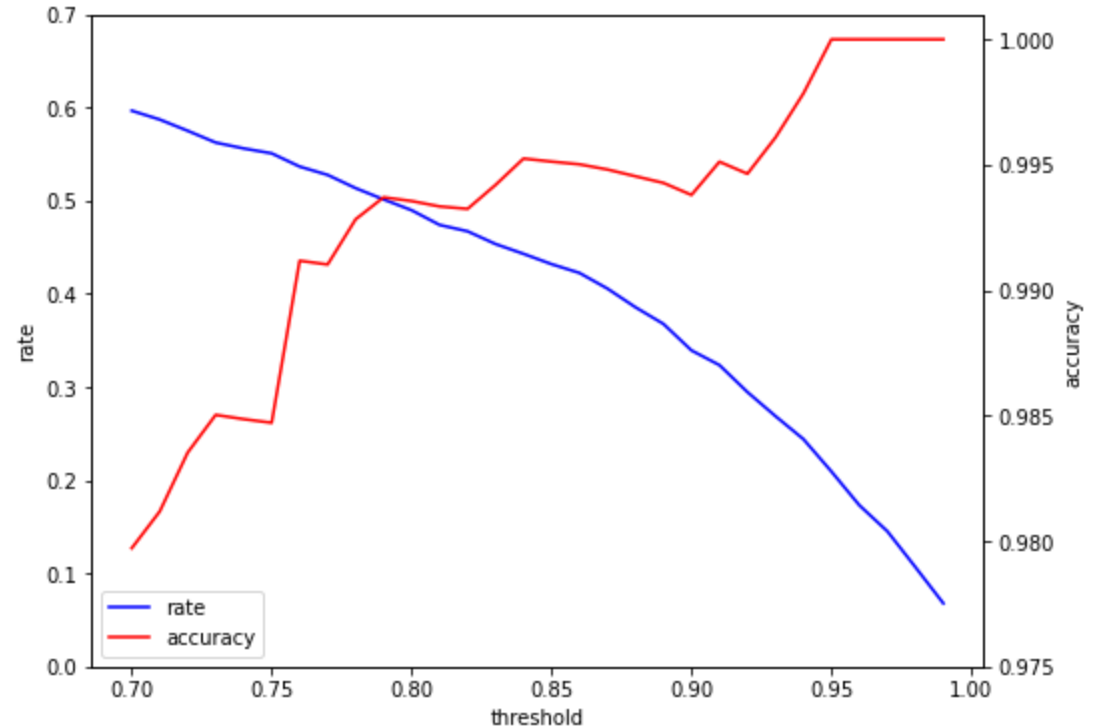
2분류로 새로운 모델을 만들었을 때 3~4점의 F1 score 상승을 확인할 수 있었다.
특히 분류 임계값을 조절할 경우 높은 정확도로 조사를 분류할 수 있었다.

분류 정확도 향상

2. 이분류 모델링

자동지급+심사	조사	largest	name
0.376	0.624	0.624	조사
0.352	0.648	0.648	조사
0.048	0.952	0.952	조사
0.008	0.992	0.992	조사
0.000	1.000	1.000	조사
...
0.050	0.950	0.950	조사
0.100	0.900	0.900	조사
0.266	0.734	0.734	조사
0.222	0.778	0.778	조사
0.496	0.504	0.504	조사

→ 일정 확률 이상만 추출



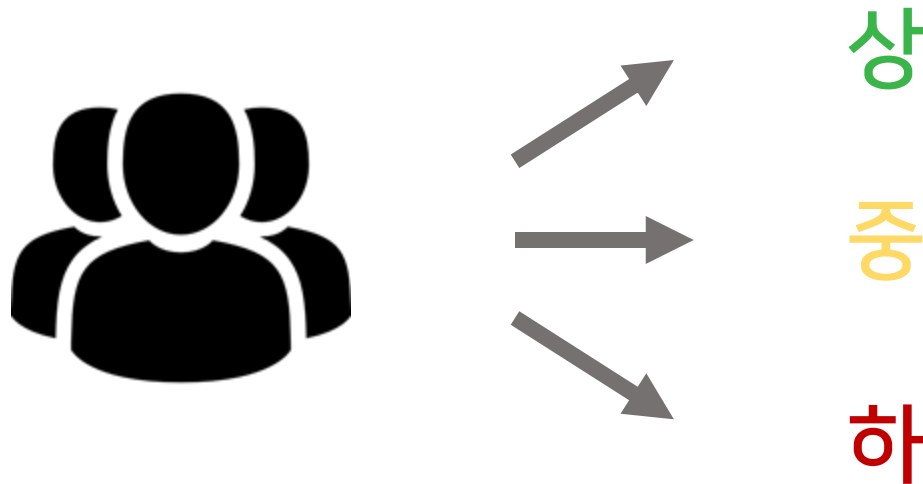
2분류 모델 예측 결과에서 조사로 예측한 확률이 0.84 이상인 예측 데이터만을 추출했을 때
총 1901개의 조사 건 중 약 45%의 데이터를 99.5%의 확률로 예측하였다.

조사 예측 확률 0.95 이상인 데이터만을 추출했을 때는 조사 건 중 21% 데이터를 100% 확률로 예측했다.

개인화 및 고객 서비스

1. 개인화 서비스

고객에게 청구예측결과를 직접 알려줄 경우 악용될 여지가 있다.
따라서 고객에게 개별 청구건에 대한 예측 결과를 직접 알려 주지 않으면서
선의의 고객에게는 혜택을, 악의의 고객에게는 패널티를 주는 방식이 무엇이 있을까?



고객이 본인의 등급을 확인하게 하여 스스로 자신의 등급을 관리하게 하며
높은 등급에게는 혜택을, 낮은 등급에게는 혜택이 없거나 패널티를 주는 방식으로
회사가 본인을 주목하고 있다는 압박을 줄 수 있다.

개인화 및 고객 서비스

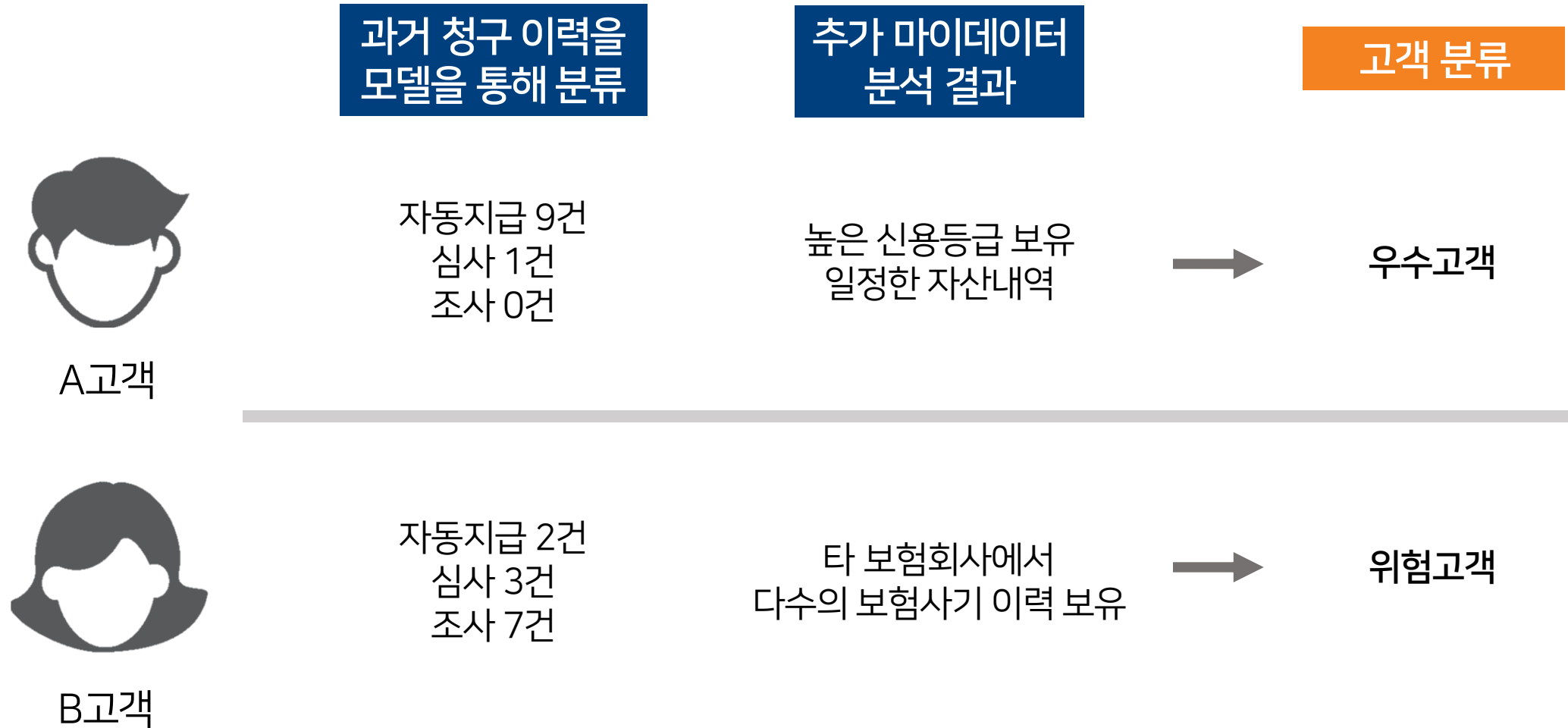
1. 개인화 서비스



마이데이터를 활용한다면 고객의 통합 금융 조회 가능하기 때문에, **고객의 연소득, 자산, 주 소비 및 취미 패턴(카드정보로 가능), 질병/입원 기록 등을 종합 고려**하여 서비스의 정확도를 향상시키는 데에 사용할 수 있다.

개인화 및 고객 서비스

1. 개인화 서비스



개인화 및 고객 서비스

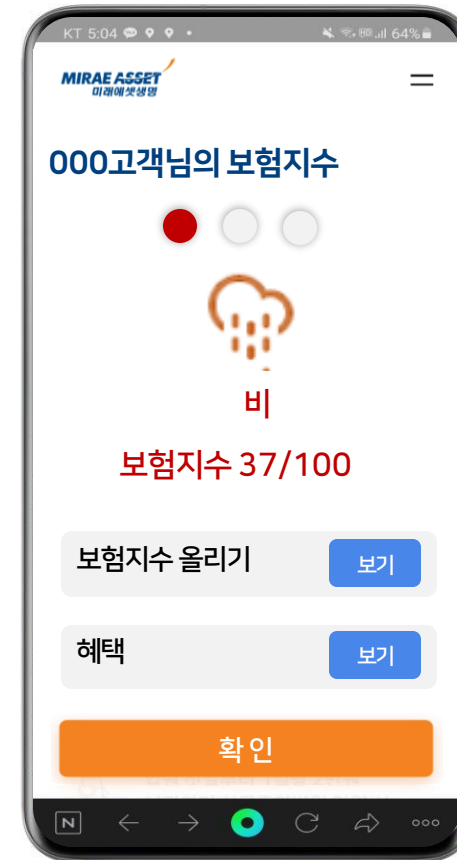
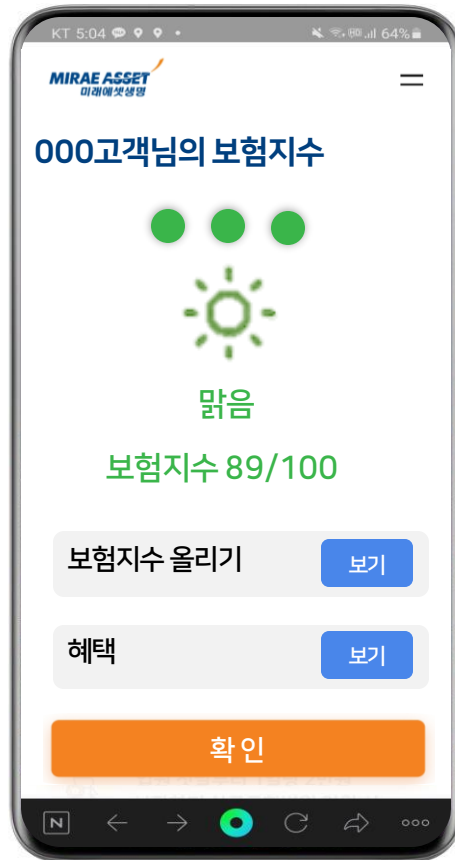
1. 개인화 서비스(예시)

서론

데이터 해석

모델링

비즈니스 활용방안

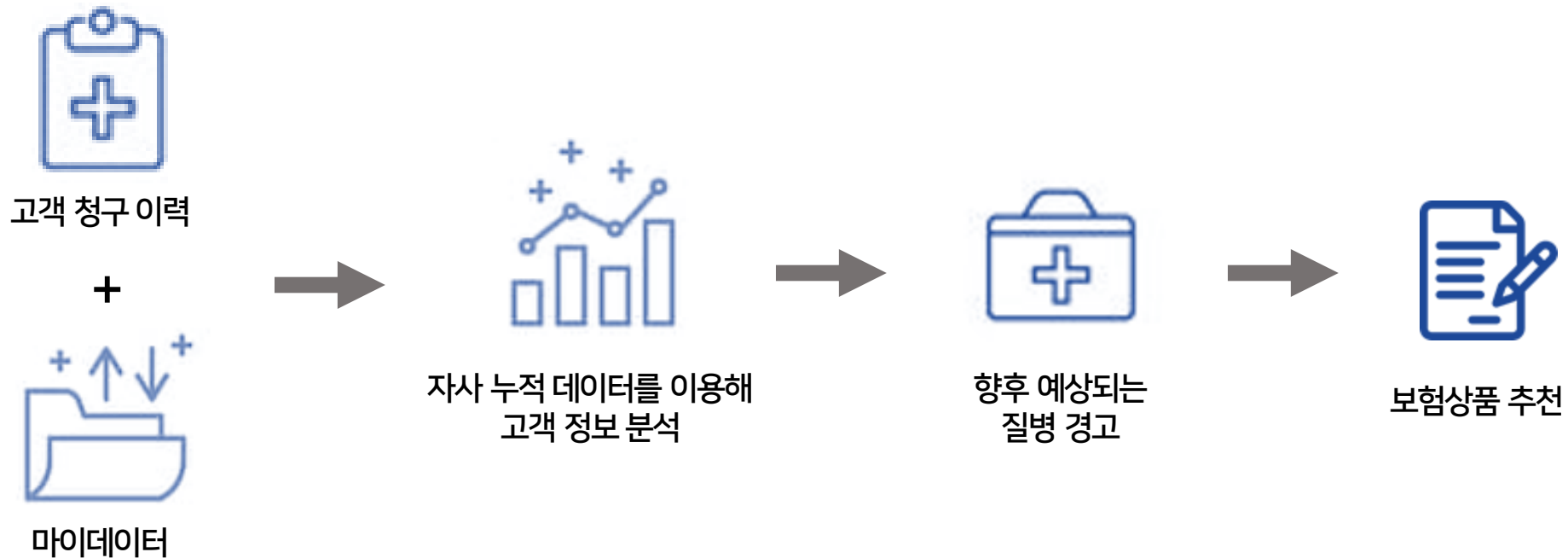


서비스 예시

개인화 및 고객 서비스

2. 헬스케어 및 보험추천 서비스

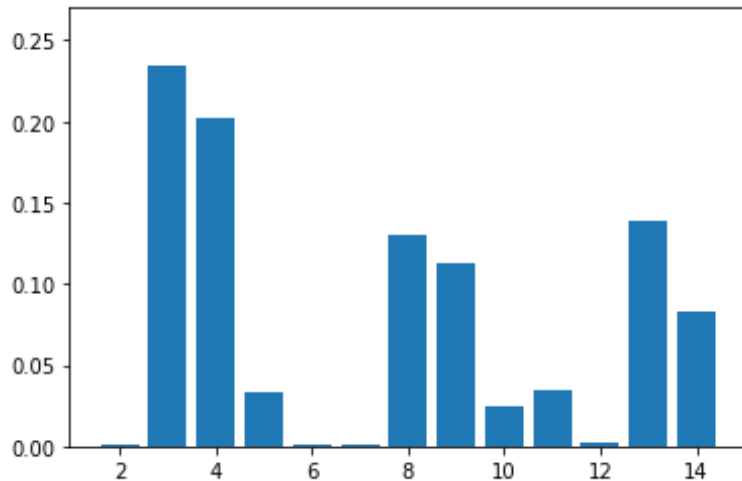
고객별 청구 내역을 알 수 있다면 이를 기존 고객들의 누적 데이터를 활용해
향후 어느 질병에 걸릴 가능성이 높고, 또 얼마 정도의 보험금을 청구할 지 예측할 수 있을 것이다.
이를 통해 고객에게 어떤 중증 질환에 걸릴 수 있는지를 경고하고, 해당 질병을 보장하는
보험상품을 추천함으로써 보험상품 가입을 유도할 수 있다.



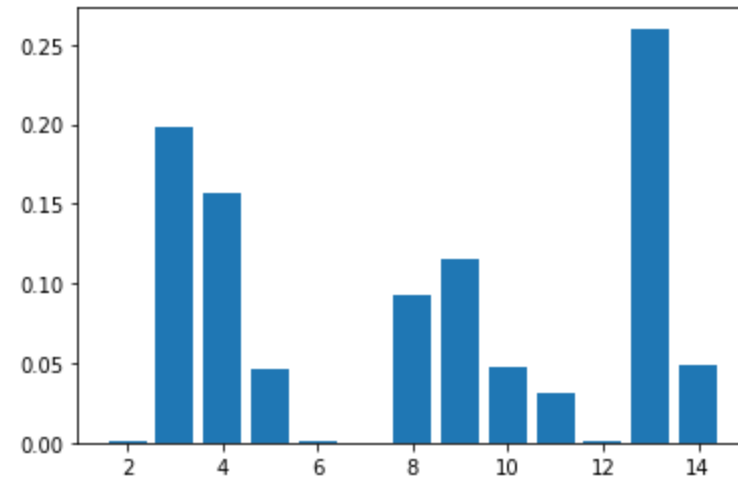
개인화 및 고객 서비스

2. 헬스케어 및 보험추천 서비스

질병 예측 시스템을 이용하여 보험상품을 추천할 때 추가적으로 대회 데이터를 활용할 수도 있다.
 대회 데이터를 분석해 보았을 때 같은 연령대라도 지역에 따라 질병비율이 달랐다.
 이러한 데이터 특성을 추가하면 좀 더 정확한 보험상품 추천이 가능할 것이다.



전국 40대 고객의 질병 분포



광주,전남,전북 40대 고객의 질병 분포

개인화 및 고객 서비스

2. 헬스케어 및 보험추천 서비스



A고객 : 남성/40대/전라남도 거주

Q. 보험상품이 너무 많아서 어떤 걸 가입해야 할지 모르겠어요. 저에게 맞는 상품을 추천해 주실 수 있나요?

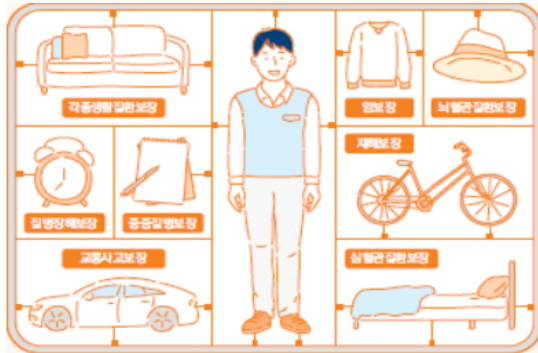


A. 고객님의 청구 내역 분석 결과 고객님의 향후 a 질병에 걸릴 확률이 높습니다.(타 고객 대비 25% ↑)
특히 거주하시는 지역의 경우, 타 지역 대비 근골격계 질환의 청구 비율이 상대적으로 높았습니다.
(전국 14%, 전남 27%)
따라서 a 질병을 보장하는 A 상품, 혹은 근골격계 질환을 특약으로 보장하는 B상품을 추천합니다.

개인화 및 고객 서비스

2. 헬스케어 및 보험추천 서비스

내가 설계하는 보장보험 (무)202005

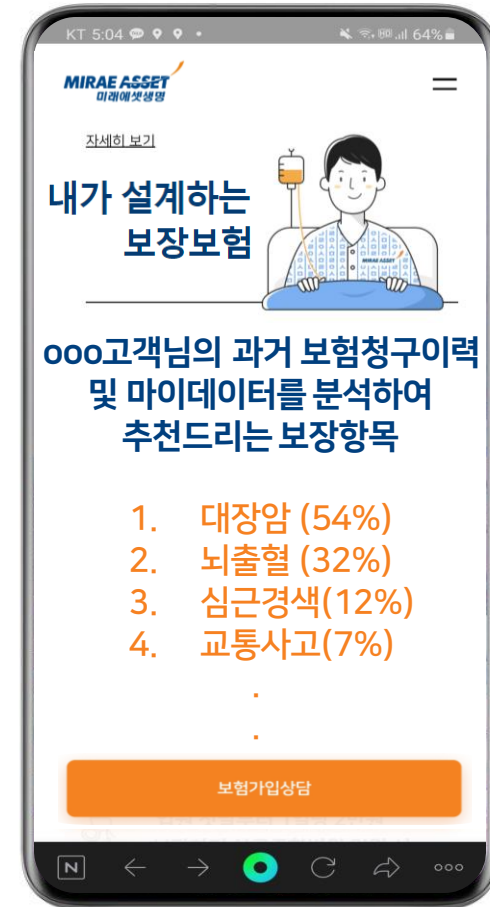


준법감시인심의필-20-06-026(2020.06.15)

원하는 보장만 편리하게 조립하고
각종 위험을 든든하게 보장받는
내가 설계하는 보장보험 (무)202005

가입시 유의사항 ▶

또한 이러한 예측결과를 '내가 설계하는 보장보험' 서비스에 접목하여 고객에게 향후 예상되는 질병과 그와 관련된 보장항목을 순위별로 나열해 선택을 도와주는 서비스도 실현 가능하다.



서비스 예시

감사합니다

Thank You!

