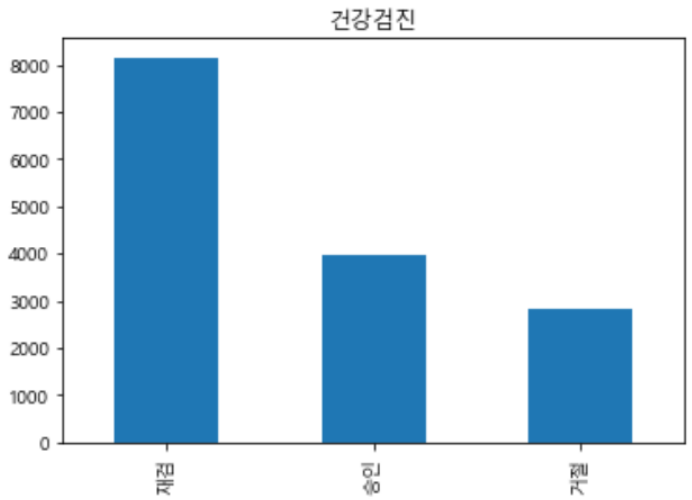
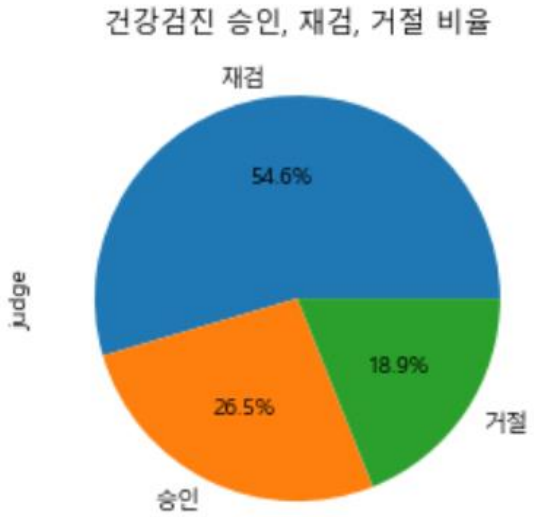
보험사 수익 극대화를 위한 가입 거절 고객에 대한 분석

A반 2조 김효진

POSCO AI · Big Data 아카데미

CONTENTS

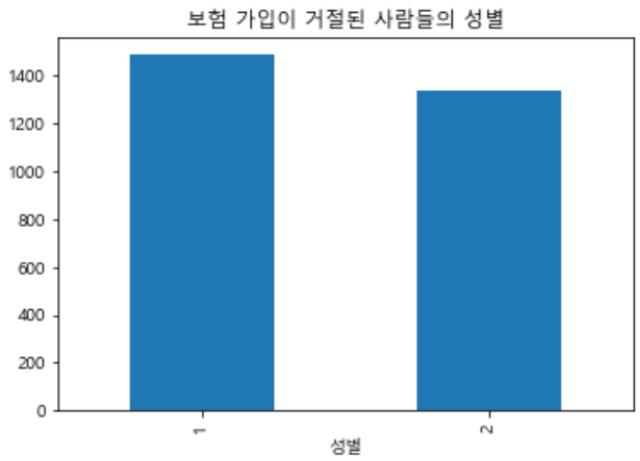


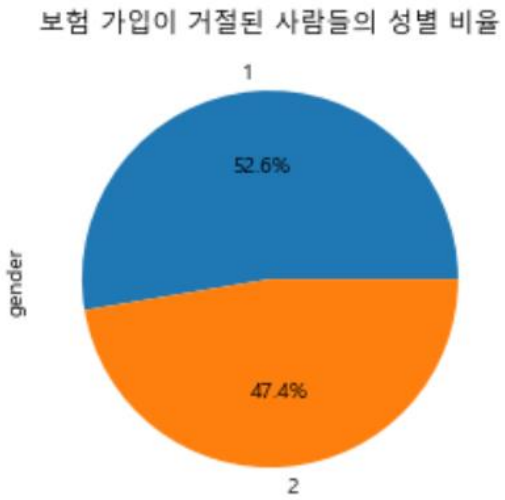


건강검진을 받은 사람 수:14938명 (insu_pre_review.csv)

- 재검: 8149명 (54.6%)
- 승인: 3962명 (26.5%)
- 거절: 2827명 (18.9%)

- 전체 건강검진 자 중 18.9%인 2827명이 가입 거절됨

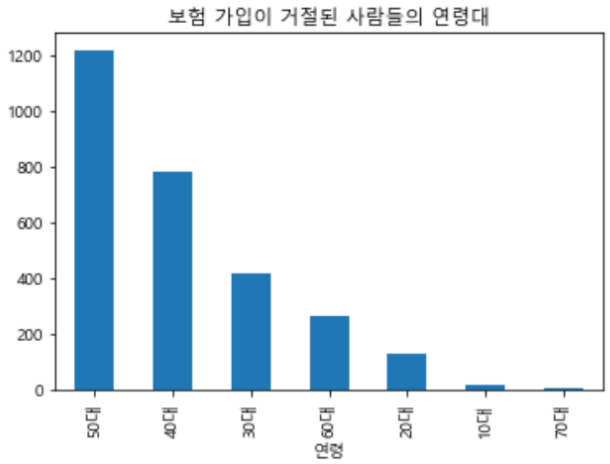


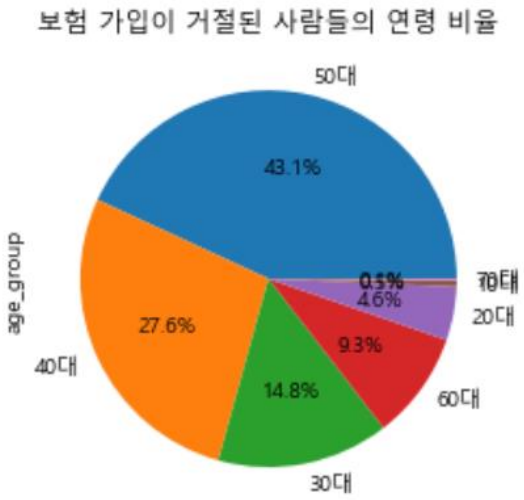


건강검진을 받은 사람 중 거절된 사람 수: 2827명 (insu_pre_review.csv)

- 남성(1): 1488명 (52.6%)
- 여성(2): 1339명 (47.4%)

- 보험 가입이 거절된 사람들의 성비는 거의 균등

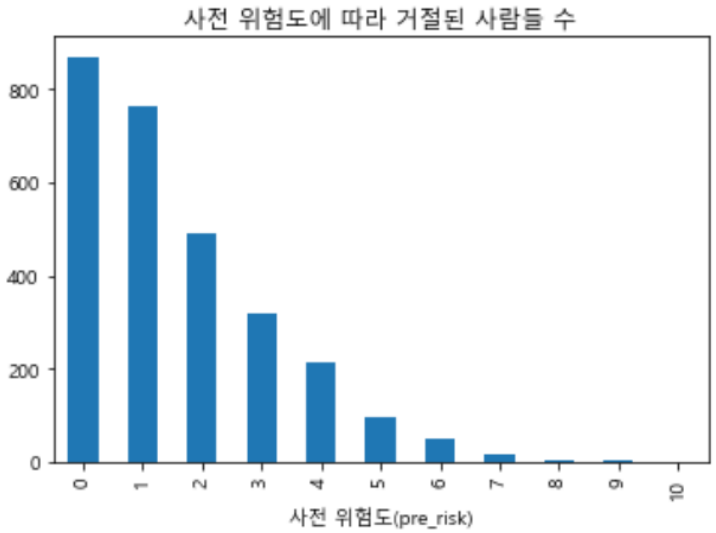


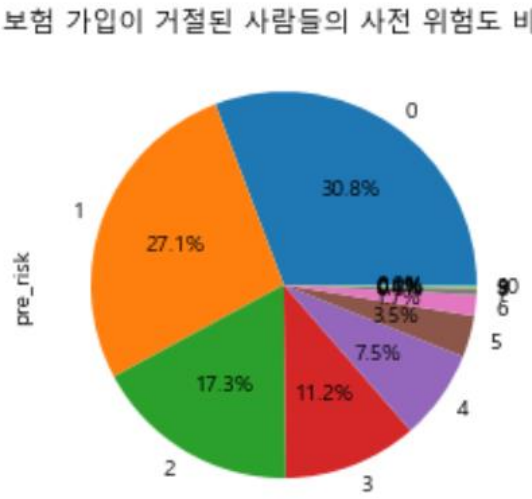


건강검진을 받은 사람 중 거절된 사람 수: 2827명 (insu_pre_review.csv)

- 10대: 15명 (0.5%)
- 20대: 130명 (4.6%)
- 30대: 417명 (14.8%)
- 40대: 780명 (27.6%)
- 50대: 1219명 (43.1%)
- 60대: 262명 (9.3%)
- 70대: 4명 (0.1%)

- 보험 가입이 거절된 사람들 중 50대가 가장 많은 비중을 차지하며 40대가 그 뒤를 이음





※ 사전 위험도(pre_risk): 혈액검사 판정결과 점수 합(judge_score) + 혈압 판정결과(bp_judge)
건강검진으로 고객 위험의 사전적인 위험을 측정하기 위해 만든 파생변수
※ 사전위험도 값이 클수록 혈압과 혈액검사에서 이상이 많이 발견된 것

```
# 보험 가입 거절된 사람들의 사전 위험도
df_count_reject_pre_risk = df_reject['pre_risk'].value_counts()
df_count_reject_pre_risk
```

```
0     870
1     766
2     489
3     318
4     213
5      98
6      49
7      16
8       4
9       3
10      1
Name: pre_risk, dtype: int64
```

가입 거절된 사람들 중 사전 위험도가 0 인 사람들이 870명
→ 혈압과 혈액검사에서 전혀 이상이 없었음에도 가입 거절된 경우로 혈압과 혈액검사가 위험 고객을 적절히 분류할 수 있음이 밝혀진다면 **해당 고객은 위험도가 낮은 고객이므로 유치할 필요**가 있음

sum_req_amount	avg_pay_amount	avg_cum_amount	dif_req_pay	risk
424498	724498	1320000	-300000	1
198600	490079	1800000	-291479	1
808750	1008750	600000	-200000	1
333680	553680	240000	-220000	1
412210	412210	375000	0	0

※ 청구 – 지급(dif_req_pay):
고객의 청구금액(sum_req_amount) – 보험사의 지불금액(avg_pay_amount)
※ 위험도(risk):
· 보험사의 수익성을 측정할 수 있는 지표
· 청구 – 지급(dif_req_pay) >= 0 일 경우 0 (보험사의 수익성 개선)
· 청구 – 지급(dif_req_pay) < 0 일 경우 1 (보험사의 수익성 악화)

```
# 설명변수
df_blood_x.columns
```

```
Index(['judge_score', 'bp_judge', 'pulse_count_judge', 'bt_chol_judge',
       'bt_crea_judge', 'bt_gluc_judge', 'bt_hb_judge', 'bt_hbsa_judge',
       'bt_hct_judge', 'bt_mch_judge', 'bt_mchc_judge', 'bt_mvc_judge',
       'bt_plat_judge', 'bt_rbc_judge', 'bt_wbc_judge', 'bt_rgpt_judge',
       'bt_sgot_judge', 'bt_sgpt_judge', 'bt_trig_judge'],
      dtype='object')
```

- 목표변수: 위험도(risk)
- 설명변수: 판정결과 합(judge_score), 혈압 판정결과(bp_judge), 맥박 판정결과(pulse_count_judge), 콜레스테롤 판정결과(bp_chol_judge) 등 총 19개 변수
- 데이터 셋: BHHJ_risk_data.csv (insu_request와 insu_pre_review를 합쳐 청구번호(req_id)별로 정리한 데이터 셋)
- 분석 데이터: 혈액검사자의 청구 데이터 14769개

	Train	Test
의사결정나무	0.788	0.785
랜덤 포레스트	0.796	0.783
그래디언트 부스팅	0.792	0.786

분석결과:

- 전체 청구데이터 수: 19115 (BHHJ_risk_data.csv)
- 혈액검사자 데이터 수: 14769
- Train: 0.4, Validation: 0.3, Test: 0.3 비율
- 의사결정나무, 랜덤 포레스트, 그래디언트 부스팅 분류 모델 사용
- Train, Test 데이터 셋의 설명력이 0.8 이하로 낮은 편

제외질병 코드

- F : 정신 및 행동 장애
- H : 눈, 귀 관련 질환
- L : 피부관련 질환
- M : 근골결계통 질환
- O : 임신, 산후 관련 질환
- P : 출생전 후기에 특정 병태
- Q : 선천기형, 염색체 이상
- R : 분류되지 않은 증상, 징후, 이상소견
- S,T : 손상, 외인에 의한 결과
- V/Y/Z/U : 사망의 외인 / 건강상태 / 특수목적 코드

혈액검사 질병

- 빈혈
- 백혈병
- 당뇨병
- 고지혈증
- 갑상선 관련 질병
- 간염
- 신부전
- 종양
- 면역결핍

- 분석 후 조원들과 논의 결과 데이터에서 **혈압 및 혈액검사로 분류하기 어려운 질병들 다수 발견**
  ex) S02: 두개골 골절, S13: 목부위의 관절 및 인대의 탈구, 염좌 및 긴장
- 따라서 **혈압 및 혈액검사로 분석 가능한 데이터들을 선별하여 새롭게 분석**
  해당 질병: 빈혈, 백혈병, 당뇨, 고지혈증, 갑상선 관련 질병, 간염, 신부전, 종양, 면역결핍
  (출처: 국민건강지식센터)
  (질병코드 A/B/C/D/E/G/I/J/K/N)
- 또한 **사전 위험도(pre_risk)가 0임에도 위험도(risk)가 1**이 나온 경우 혈압 및 혈액검사로 예측하기 어려운 질병들에 걸린 것임을 확인하여 분석에서 제외
  ex) K63: 장의 기타 질환, N20: 신장 및 요관의 결석, J15: 세균성 폐렴

19100	C134221-20120214-16711-01-001	C134221	혈액검사	재검	2	51	입원	19	건강보험	K21	17	456361
19103	C134221-20130304-12943-04-001	C134221	혈액검사	재검	2	51	입원	12	건강보험	I84	30	298363
19105	C134221-20130917-12901-02-001	C134221	혈액검사	재검	2	51	입원	23	건강보험	J02	36	558396
19106	C134227-20100423-13218-01-001	C134227	혈액검사	재검	2	45	외래	1	건강보험	J00	5	42600
19107	C134227-20100423-13243-01-001	C134227	혈액검사	재검	2	45	외래	1	건강보험	J06	5	11120

5031 rows × 61 columns

※ 청구 – 지급(dif_req_pay):
고객의 청구금액(sum_req_amount) – 보험사의 지불금액(avg_pay_amount)
※ 위험도(risk):
· 보험사의 수익성을 측정할 수 있는 지표
· 청구 – 지급(dif_req_pay) >= 0 일 경우 0 (보험사의 수익성 개선)
· 청구 – 지급(dif_req_pay) < 0 일 경우 1 (보험사의 수익성 악화)

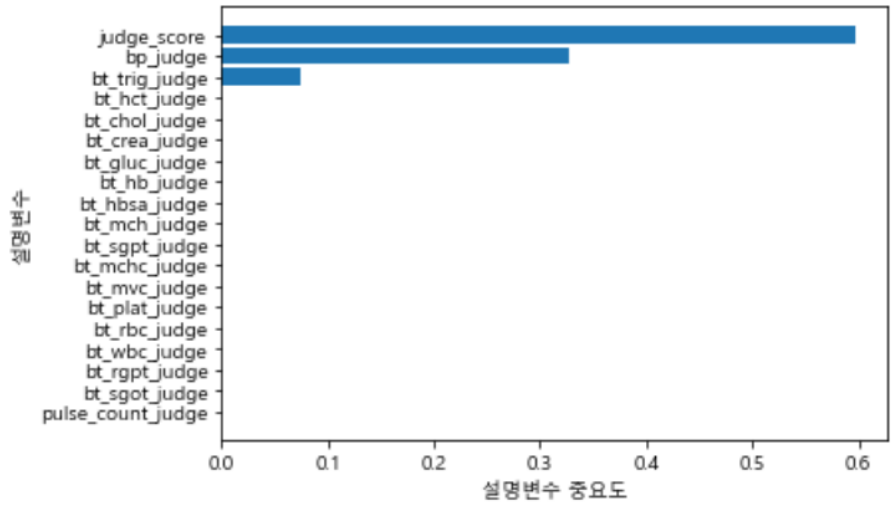
```
# 설명변수
df_blood_x.columns
```

```
Index(['judge_score', 'bp_judge', 'pulse_count_judge', 'bt_chol_judge',
       'bt_crea_judge', 'bt_gluc_judge', 'bt_hb_judge', 'bt_hbsa_judge',
       'bt_hct_judge', 'bt_mch_judge', 'bt_mchc_judge', 'bt_mvc_judge',
       'bt_plat_judge', 'bt_rbc_judge', 'bt_wbc_judge', 'bt_rgpt_judge',
       'bt_sgot_judge', 'bt_sgpt_judge', 'bt_trig_judge'],
      dtype='object')
```

- 목표변수: 위험도(risk)
- 설명변수: 판정결과 합(judge_score), 혈압 판정결과(bp_judge), 맥박 판정결과(pulse_count_judge), 콜레스테롤 판정결과(bp_chol_judge) 등 총 19개 변수
- 데이터 셋: BHHJ_risk_data.csv (insu_request와 insu_pre_review를 합쳐 청구번호(req_id)별로 정리한 데이터 셋)
- 분석 데이터: 질병코드 A/B/C/D/E/G/I/J/K/N을 주상병으로 가지며 사전 위험도(pre_risk)가 0이고 위험도(risk)가 1인 데이터를 제외한 혈액검사자 데이터 수: 5031개

	Feature	Importance
0	judge_score	0.598
1	bp_judge	0.328
18	bt_trig_judge	0.074
3	bt_chol_judge	0.000
4	bt_crea_judge	0.000
5	bt_gluc_judge	0.000
6	bt_hb_judge	0.000
7	bt_hbsa_judge	0.000
8	bt_hct_judge	0.000
2	pulse_count_judge	0.000
10	bt_mchc_judge	0.000
11	bt_mvc_judge	0.000
12	bt_plat_judge	0.000
13	bt_rbc_judge	0.000
14	bt_wbc_judge	0.000
15	bt_rgpt_judge	0.000
16	bt_sgot_judge	0.000
17	bt_sgpt_judge	0.000
9	bt_mch_judge	0.000

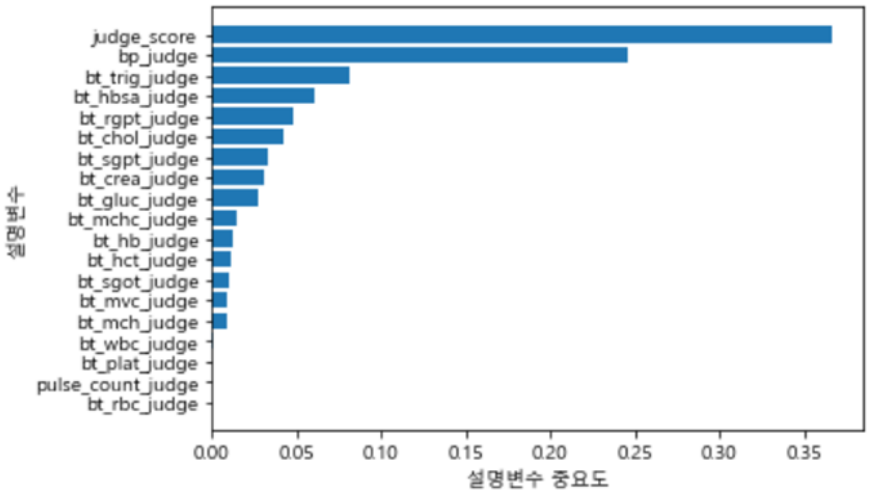
```
Accuracy on training set:0.937
Accuracy on test set:0.937

Confusion matrix:
[[1415    0]
 [  95    0]]
```



- 결과: 의사결정나무 최종 모델
  - 최대 깊이: 10
  - 분리 노드: 15
  - 최소 잎사귀 수: 5
- Train 데이터 설명력: 0.937
- Test 데이터 설명력: 0.937
- 분석 결과 **혈액검사 판정결과의 합**(judge_score)가 가장 중요한 변수로 파악되며 두번째와 세번째로 **혈압 판정결과**(bp_judge)와 **중성지방 판정결과**(bt_trig_judge)이 중요하다고 판단됨

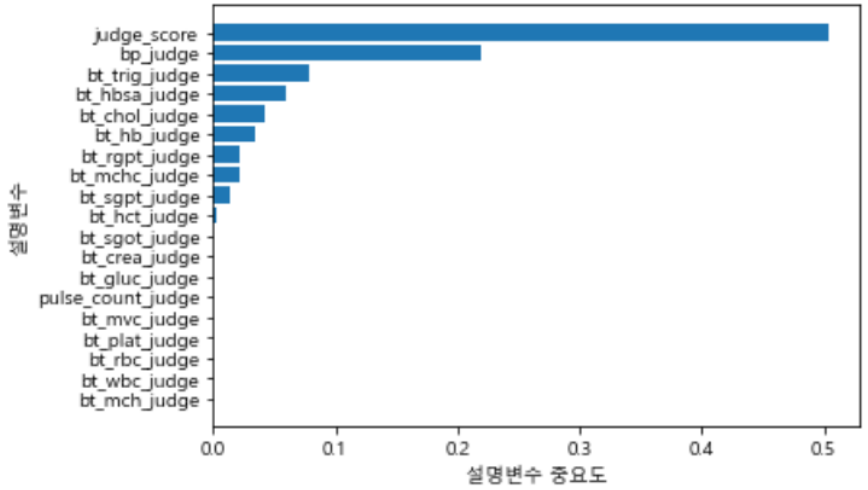
	Feature	Importance
0	judge_score	0.348
1	bp_judge	0.240
18	bt_trig_judge	0.094
7	bt_hbsa_judge	0.068
17	bt_sgpt_judge	0.050
15	bt_rgpt_judge	0.042
3	bt_chol_judge	0.041
4	bt_crea_judge	0.029
5	bt_gluc_judge	0.025
10	bt_mchc_judge	0.019
6	bt_hb_judge	0.012
8	bt_hct_judge	0.011
11	bt_mvc_judge	0.008
16	bt_sgot_judge	0.007
9	bt_mch_judge	0.005
14	bt_wbc_judge	0.001
2	pulse_count_judge	0.000
12	bt_plat_judge	0.000
13	bt_rbc_judge	0.000

```
Accuracy on training set:0.937
Accuracy on test set:0.937

Confusion matrix:
[[1415    0]
 [  95    0]]
```



- 결과: 랜덤 포레스트 최종 모델
  - 최대 깊이: 7
  - 분리 노드: 4
  - 최소 잎사귀 수: 10
  - 나무 수: 150
- Train 데이터 설명력: 0.937
- Test 데이터 설명력: 0.937
- 분석 결과 **혈액검사 판정결과의 합**(judge_score)가 가장 중요한 변수로 파악되며 두번째와 세번째로 **혈압 판정결과**(bp_judge)와 **중성지방 판정결과**(bt_trig_judge)이 중요하다고 판단됨

	Feature	Importance
0	judge_score	0.504
1	bp_judge	0.219
18	bt_trig_judge	0.079
7	bt_hbsa_judge	0.060
3	bt_chol_judge	0.042
6	bt_hb_judge	0.035
15	bt_rgpt_judge	0.022
10	bt_mchc_judge	0.022
17	bt_sgpt_judge	0.014
8	bt_hct_judge	0.003
4	bt_crea_judge	0.000
5	bt_gluc_judge	0.000
2	pulse_count_judge	0.000
11	bt_mvc_judge	0.000
12	bt_plat_judge	0.000
13	bt_rbc_judge	0.000
14	bt_wbc_judge	0.000
16	bt_sgot_judge	0.000
9	bt_mch_judge	0.000



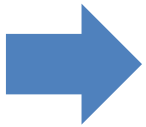

```
Accuracy on training set:0.941
Accuracy on test set:0.934

Confusion matrix:
[[1408    7]
 [  93    2]]
```

- 결과: 그래디언트 부스팅 최종 모델
  - 최대 깊이: 10
  - 분리 노드: 15
  - 최소 잎사귀 수: 5
- Train 데이터 설명력: 0.941
- Test 데이터 설명력: 0.934
- 분석 결과 **혈액검사 판정결과의 합**(judge_score)가 가장 중요한 변수로 파악되며 두번째와 세번째로 **혈압 판정결과**(bp_judge)와 **중성지방 판정결과**(bt_trig_judge)이 중요하다고 판단됨

	Train	Test
의사결정나무	0.937	0.937
랜덤 포레스트	0.937	0.937
그래디언트 부스팅	0.941	0.934

분석결과:
- 전체 청구데이터 수: 19115 (BHHJ_risk_data.csv)
- 질병코드 A/B/C/D/E/G/I/J/K/N을 주상병으로 가지며 사전 위험도(pre_risk)가 0이고 위험도(risk)가 1인 데이터를 제외한 혈액검사자 데이터 수: 5031
- Train: 0.4, Validation: 0.3, Test: 0.3 비율
- 의사결정나무, 랜덤 포레스트, 그래디언트 부스팅 분류 모델 사용
- Train, Test 데이터 셋의 설명력이 0.93 이상으로 높은 편
- 변수 중요도는 세 모델 모두 **혈액검사 판정결과 합(judge_score), 혈압 판정결과(bp_judge), 중성지방 판정결과(bp_trig_judge) 순**으로 중요성을 보임

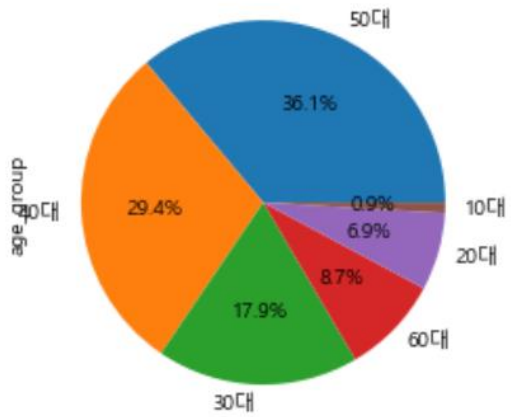
데이터 수: 14769
(모든 혈액검사자)
목표변수: 위험도(risk)
설명변수(혈압 및 혈액검사의 판정결과)
설명력: 0.8 이하로 낮은 편



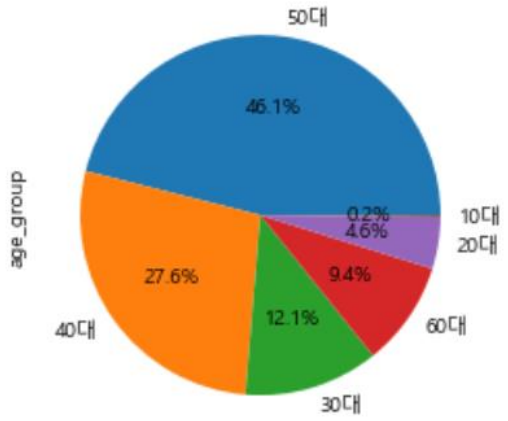
데이터 수: 5031
(혈액검사로 예측 및 분류 가능한 질병을 가진 혈액검사자)
목표변수: 위험도(risk)
설명변수(혈압 및 혈액검사의 판정결과)
설명력: 0.93 이상으로 높은 편

→ 혈압 및 혈액검사에 의해 위험 고객을 적절히 분류할 수 있음이 밝혀짐
→ 혈압 및 혈액검사로 예측 및 분류 가능한 데이터(5031)와 보험 가입이 거절된 사람들의 데이터 특성을 비교한 후 적절한 보험 상품을 추천하여 수익성 향상 가능

사전 위험도(pre_risk)가 0이며 보험 가입이 거절된 사람들의 연령 비율



사전 위험도(pre_risk)가 0 이며 보험 가입이 승인된 사람들의 연령 비율



- 사전 위험도가 0이며 보험 가입이 거절된 사람(870명)과 사전 위험도가 0이며 보험 가입이 승인된 사람(2310명)의 연령 분포는 비슷함(1위: 50대, 2위: 40대 3위: 30대)
- 따라서 **연령대별 보험 수익성을 분석**하여 보험 가입이 거절된 사람에게 적절한 보험 상품을 추천할 필요가 있음

조심조심 보험 4	749
건강보험	585
울라트 보험	315
평생 건강 보장 1	260
All My Life 2	154
건강 보살핌	65
All My Life 1	34
가족 만족 보험 2+	28
안심보험	26
통합보험 +2	22
건강+행복 보험	18
통합보험 +3	16
단체보험(상해)	14
조심조심 보험 1	10
통합보험 +1	4
가족 만족 보험 1	3
평생 보험 2	3
가족 만족 보험 2	2
건강보험 3	1
평생 보험 1	1

※ 사전 위험도가 0인 보험 가입자의 보험 종류별 가입 수

조심조심 보험 4	10대	0.0
	20대	0.0
	30대	2079.0
	40대	3082.0
	50대	7101.0
	60대	1247.0
건강보험	20대	0.0
	30대	9344.0
	40대	270.0
	50대	11446.0
	60대	7070.0
울라트 보험	20대	0.0
	30대	3082.0
	40대	0.0
	50대	42.0
	60대	0.0
평생 건강 보장 1	10대	0.0
	30대	7562.0
	40대	24204.0
	50대	23761.0
	60대	0.0
All My Life 2	30대	47299.0
	40대	48791.0
	50대	87868.0
	60대	35556.0

※ 보험상품을 기준으로 한 연령대별 청구 – 지급(dif_req_pay) 금액의 평균

- 사전 위험도가 0인 사람들 중 보험 가입이 승인된 사람(2310명)은 위와 같이 조심조심 보험 4(749명), 건강보험(585명), 울라트 보험(315명) 등에 많이 가입
- 사전 위험도가 0이며 보험 가입이 거절된 사람(870명)들이 **50대, 40대, 30대의 순으로 많은** 것을 고려하면 **All My Life 2 보험**을 추천하여 이들에게 가입하도록 유도하는 것이 보험사의 수익성 극대화를 위해 가장 좋은 방법이라 생각됨