

B i g d a t a P r o j e c t

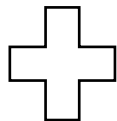
A 반 2 조 강 지 영



데이터 구성

보험 청구 지급 정보 (insu_request.csv)

- **고객 ID**
- 성별
- 연령
- 검사 구분(혈액/일반)
- 판정 결과(승인/재검)
- 보험 청구 금액
- 보험 지급 금액
- 누적 납입 보험료
- 보험 상품 가입 기간
- 보험 상품
- 주 상병
- 등등



보험 가입 사전 승인 검진 정보 (insu_pre_review.csv)

- **고객 ID**
- 성별
- 연령
- 검사 구분(혈액/일반)
- 신장/체중/가슴 둘레/허리 둘레/BMI/WHTR 등등
- 혈압 / 맥박
- 항목별 혈액검사 수치
(ex. 콜레스테롤, 혈당 등등)
- 항목별 혈액검사 수치에 따른 정상/비정상 판정 결과 등등

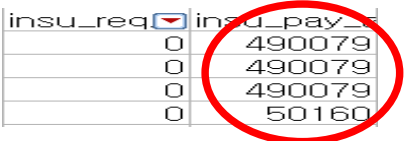


개인별 보험 청구 지급 및 검진 정보 (add_profit.csv)

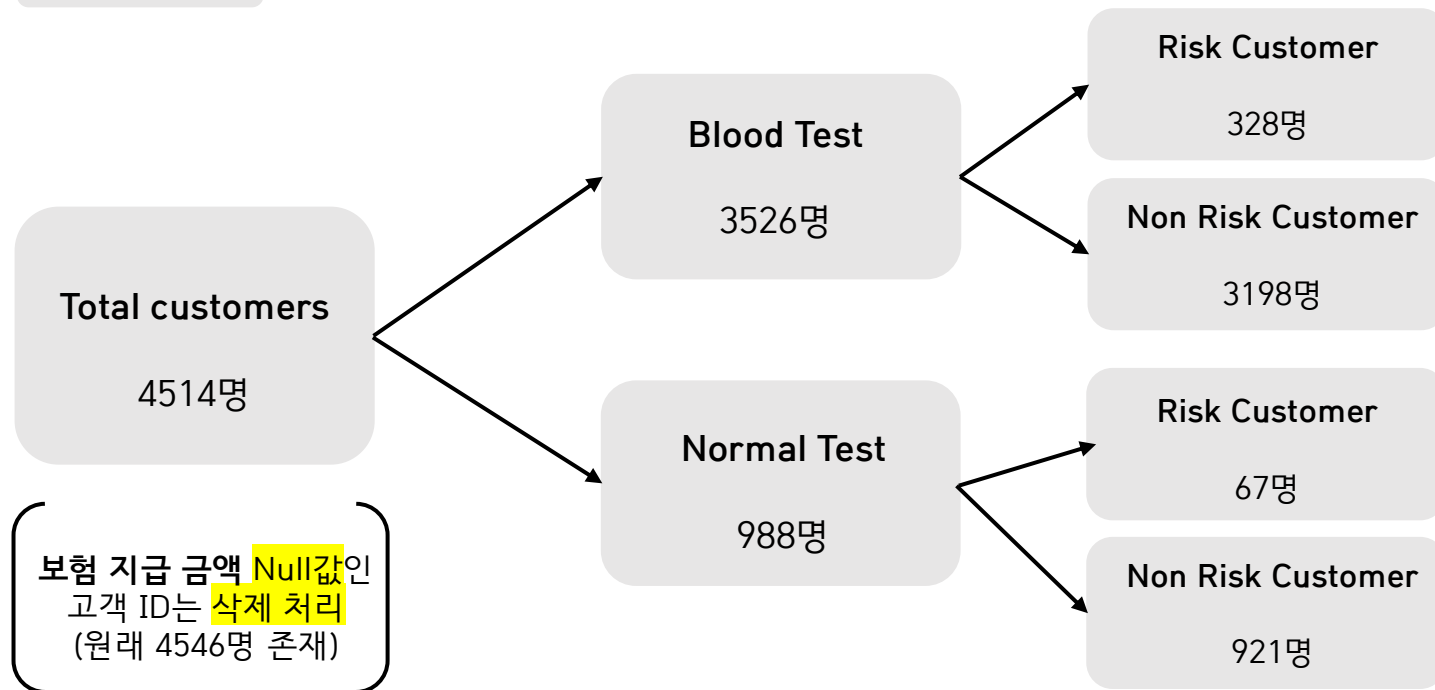
- **고객 ID** 기준으로 보험 청구 지급 정보와 보험 가입 사전 승인 검진 정보 **병합**
- 보험 청구 지급 정보와 보험 가입 승인 검진 정보에 있는 모든 변수 존재
- 수익성 판단을 위해 **"profit"** 변수 생성
- 수익성에 따른 위험/비 위험 고객으로 이진 분류한 **"risk"** 변수 생성

변수 설명

개인별 보험 청구 지급 및 검진 정보 (add_profit.csv)

| Profit (수익성) | Risk (위험 / 비 위험 고객) | 부가 설명 |
|---|---|---|
| <p>- profit : 순 수익성을 알기 위해 각 개인이 보험 회사에 납부한 총 납입액(insu_cum_amount)과 보험 회사로부터 수령한 총 지급액(insu_pay_amount)을 감산한 값</p> <p>• 개인별 총 납입액 - 총 지급액</p> | <p>- risk : 수익성을 기준으로 위험 고객/비 위험 고객 이진(binary)분류</p> <ul style="list-style-type: none"> • 납입-지급 ≥ 50000 : 비 위험 • 납입-지급 < 50000 : 위험 <p>- 0원이 아닌 50000으로 분류 기준을 선택한 이유는 보험 회사측에서 사전 검사(혈액 혹은 소변검사) 비용을 부담하기 때문에 사전 검사로 인한 손실을 충당하기 위해서는 사전 검사 비용을 기준으로 수익성 판단</p> | <p>- "납입-지급"을 선택한 이유? : 개인이 청구를 0원 하더라도 정액 보험과 같은 경우, 개인에게 지급되는 보험금이 존재</p>  <p>따라서, '청구'보다는 확실한 수익성을 계산할 수 있는 '납입'을 사용</p> |

데이터 구성



데이터 정제

[Data set : 개인별 보험 청구 지급 및 검진 정보]

- 위험 고객 여부 판별할 수 있는 것은 오직 **고객의 사전 검사 데이터**
- **개인의 특성**으로만 위험 고객 판별할 예정이므로 개인의 특성과 관련 없는 columns 제거

-보험 관련 (insu_id / insu_contract_date / insu_prod_id / insu_prod_name)

→ 사전 검사 데이터에서 각 개인이 어떤 보험을 들 것인지 알 수 없으므로 제거.

-청구 관련 (req_id / req_id_seq)

→ 청구보다 정확한 납입 데이터 사용할 예정이므로 제거.

-상병 관련 (sick_main / sick_1st / sick_2nd / sick_3rd)

→ 사전 검사 데이터만으로 개인의 병 예측 불가하므로 제거.

데이터 정제

[Data set : 개인별 보험 청구 지급 및 검진 정보]

- 위험 고객 여부 판별할 수 있는 것은 오직 **고객의 사전 검사 데이터**
- **개인의 특성**으로만 위험 고객 판별할 예정이므로 개인의 특성과 관련 없는 columns 제거

-진단 관련 (dg_cat / dg_start_date / dg_end_date / dg_duration)

→ 사전 검사만으로 미리 알 수 없는 정보이므로 제거.

dg_cat과 dg_duration은 고려 후 사용 가능성 존재.

-보험금 관련(insu_req_amount / insu_pay_amount / insu_pay_date / insu_duration / insu_cum_amount)

→ 보다 정확한 수익성을 위해 청구 보험금이 아닌 납입 보험금을 사용하기로 하였으므로 청구 금액 제거.

→ 수익성(profit)은 납입 보험금과 지급 보험금으로 파생된 변수이므로 제거.

→ 보험 기간은 납입 보험금과 상관관계가 높으므로 제거.

-검사 관련 (review_date)

→ 위험고객 판정과 검사 일자 관련 무관하므로 제거.

데이터 정제

Blood Test

| | |
|---------|------|
| bt_chol | 0 |
| bt_crea | 1058 |
| bt_gluc | 0 |
| bt_hb | 0 |
| bt_hbsa | 0 |
| bt_hct | 0 |
| bt_mch | 0 |
| bt_mchc | 0 |
| bt_mvc | 0 |
| bt_plat | 0 |
| bt_rbc | 0 |
| bt_wbc | 0 |
| bt_rgpt | 1579 |
| bt_sgot | 1492 |
| bt_sgpt | 0 |
| bt_trig | 900 |

- 혈액 검사 데이터 : 총 8개의 결측 columns 존재
(bt_crea / bt_rgpt / bt_sgot / bt_trig 와 그에 따른 판정 결과)

① 결측 데이터 있는 행 제거 or 결측 데이터 대체
: 약 1000개 이상의 데이터 손실

② 결측 데이터 존재하는 columns 제거
: 8개의 columns 손실

→ 개인별 특성 파악해서 수익성 여부를 판별하는 모델링이기 때문에,
1000개의 데이터 손실보다 8개의 columns 손실이 낫다고 판단.

→ bt_crea / bt_rgpt / bt_sgot / bt_trig 와 그에 따른 판정 결과
columns 제거 후 모델링.

- 일반 검진 데이터 : 결측 데이터 X

데이터 모델링 - 일반 검진

Decision Tree

- **Accuracy**

- Training set : 1.0

- Validation set : 0.882

- Test set : 0.923

- **Variable Importance**

- waist>age>bust>

- pulse_count>height

Random Forest

- **Accuracy**

- Training set : 0.977

- Validation set : 0.946

- Test set : 0.909

- **Variable Importance**

- Bmi>whtr>pulse_cou

- nt>age>waist

Gradient Boosting

- **Accuracy**

- Training set : 0.995

- Validation set : 0.943

- Test set : 0.909

- **Variable Importance**

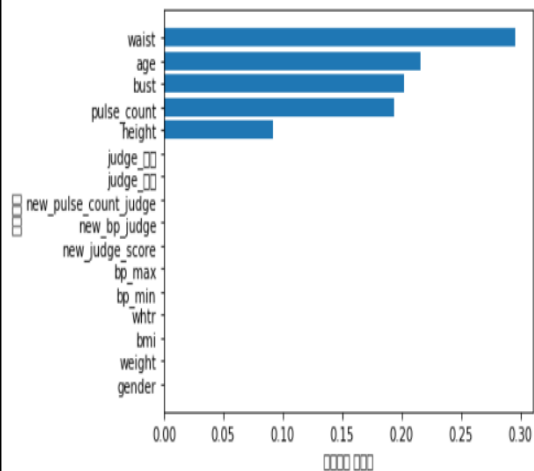
- bt_chol>bt_RBC>bt_

- hct>bt_plat>bt_gluc

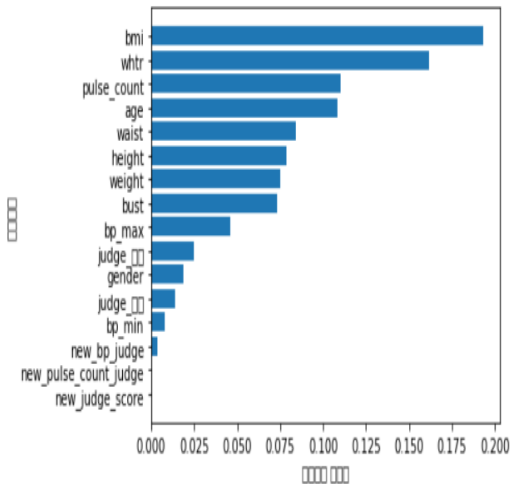
→ Decision Tree로 모델링 했을 때 가장 높은 test set 정확도를 보였다.

데이터 모델링 - 일반 검진

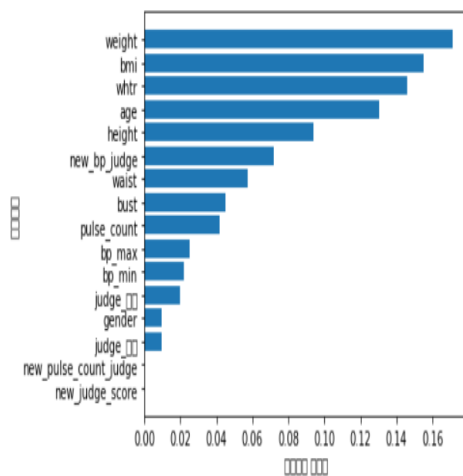
Decision Tree



Random Forest



Gradient Boosting



→ 대체적으로 age, pulse_count, height, bmi, whtr 등이 중요한 변수로 보인다.

데이터 모델링 - 혈액 검사

Decision Tree

- **Accuracy**

- Training set : 1.0
- Validation set : 0.821
- Test set : 0.923

- **Variable Importance**

Bmi (only)

Random Forest

- **Accuracy**

- Training set : 0.977
- Validation set : 0.900
- Test set : 0.923

- **Variable Importance**

Age>bt_chol>whtr>bmi>bt_plat>bt_hb

Gradient Boosting

- **Accuracy**

- Training set : 0.96
- Validation set : 0.897
- Test set : 0.923

- **Variable Importance**

Weight>bmi>whtr>age>height

→ Test set의 정확도는 모두 동일하므로, validation set의 정확도가 가장 높은 RF로 사용한다.

데이터 모델링 - 혈액 검사

Decision Tree

| Feature Importance | | |
|--------------------|-------------------------|-----|
| 5 | bmi | 1.0 |
| 0 | age | 0.0 |
| 43 | new_bt_mch_judge_0.0 | 0.0 |
| 31 | new_pulse_count_judge_0 | 0.0 |
| 32 | new_pulse_count_judge_1 | 0.0 |
| 33 | new_bt_chol_judge_0.0 | 0.0 |

Random Forest

| Feature Importance | | |
|--------------------|---------|------|
| 0 | age | 0.12 |
| 10 | bt_chol | 0.11 |
| 6 | whtr | 0.09 |
| 5 | bmi | 0.09 |
| 17 | bt_plat | 0.06 |
| 12 | bt_hb | 0.06 |
| 20 | bt_sgpt | 0.05 |
| 3 | bust | 0.05 |
| 13 | bt_hct | 0.04 |
| 1 | height | 0.04 |

Gradient Boosting

| Feature Importance | | |
|--------------------|---------|-------|
| 10 | bt_chol | 0.134 |
| 18 | bt_rbc | 0.116 |
| 13 | bt_hct | 0.087 |
| 17 | bt_plat | 0.074 |
| 11 | bt_gluc | 0.071 |
| 5 | bmi | 0.052 |
| 16 | bt_mvc | 0.051 |

→ 대체적으로 bmi, bt_chol 등이 중요한 변수로 보인다.

[보완점]

1 . 위험 고객 분류 모델 정확도 향상

- : 사전 검진에서 거절된 고객들 중 비 위험 고객 찾아서 신규 고객으로 유치할 수 있는 방안 마련.
- : 모델에 test 데이터 적용해 본 결과, Accuracy는 0.93으로 비교적 높은 수치이지만 위험 고객을 찾을 수 없으므로 개선 필요.

2 . 데이터 셋의 다양화 시도

- : Null 값을 처리하기 위해서는 데이터 셋을 쪼개서 사용할 수 밖에 없음.
- : 현재 가지고 있는 데이터 셋 제외하고, 다양한 변수의 조합으로 데이터 손실을 최소화하는 데이터 셋 생성.

[향후 계획]

- 1 . 혈액과 일반 검진 모델을 통해 위험한 고객의 특성으로 나타난 변수들을 이용
 - 수익성 향상시키는 고객과 유사한 특징을 가진 거절 고객을 신규 고객으로 유치
 - 수익성 악화시키는 고객과 유사한 특징을 가진 신규 고객에게 할증 혹은 거절

- 2 . 국민건강검진 정보 데이터 활용
 - 사전 검진 정보 데이터와 겹치는 국민 건강 검진 정보의 변수(키, 몸무게, 혈압 등등)을 이용하여, 사전 검진 정보 데이터에서 새로운 변수 생성(시력, 청력, 흡연 음주 등등) 후 모델링 정확도 향상을 시도
(ex. 키, 몸무게, 혈압 등 비슷한 사람들에게 그에 해당하는 시력, 청력, 흡연, 음주 등의 변수 값 부여)
→ 모델링에 보다 많은 변수 이용 가능