질병 예측 모델을 통한 수익성 향상 전략

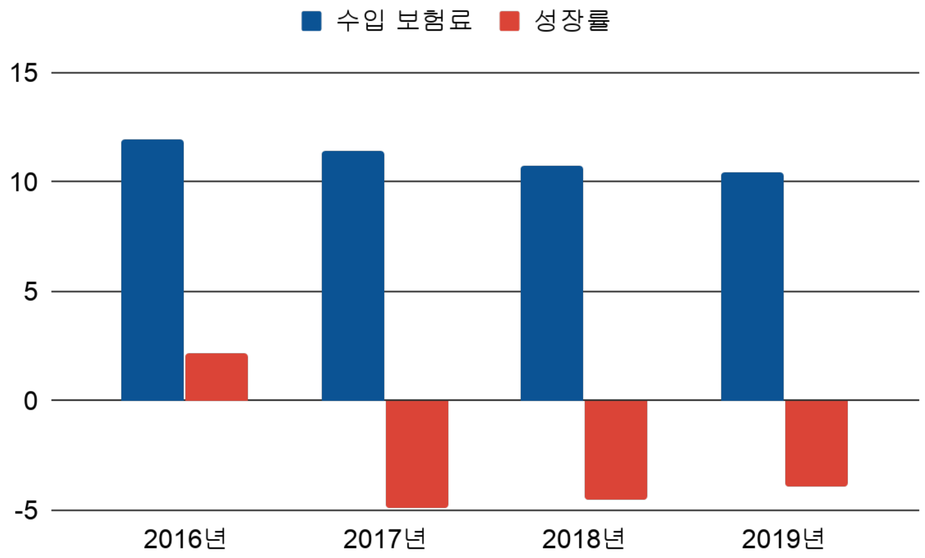
POSCO AI Big Data Academy

A반 2조 정지성

추진 배경

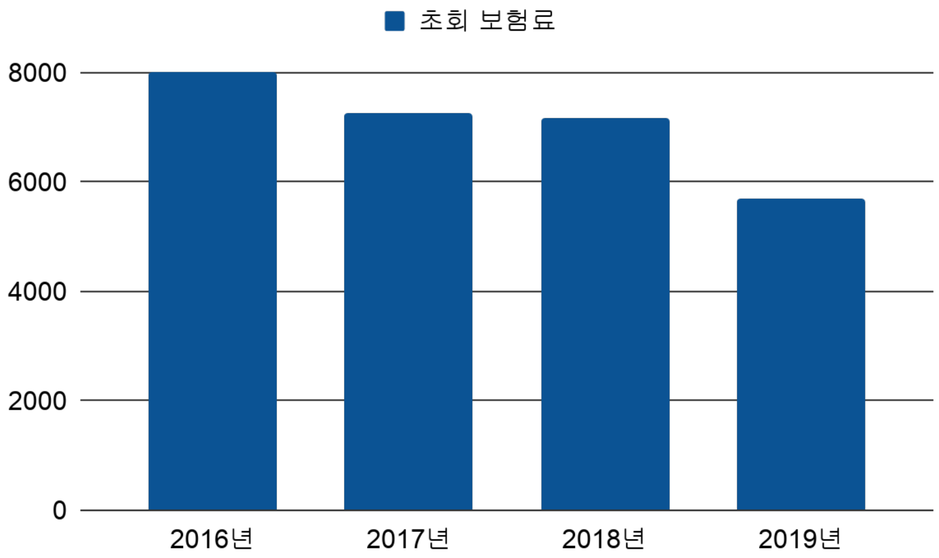
생명 보험 업계 현황

- 수입 보험료 감소
- 초회 보험료 3년 사이 40.5% 급감
- 지급보험금 증가



포빅 생명

- 2019년 초회보험료 5,690억원
- 초회 보험료 감소
- 가입 거절 비율 19.8%



추진 배경

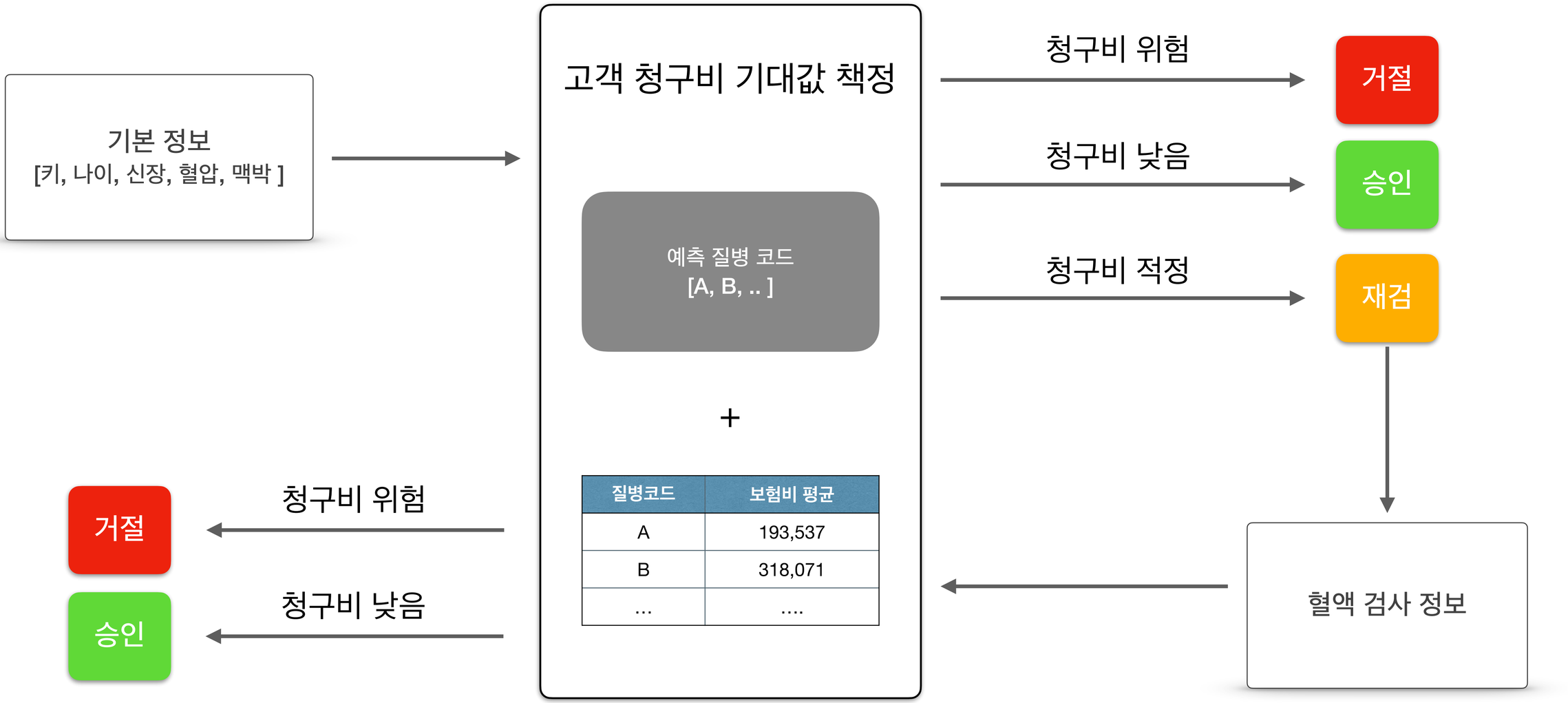
문제점 도출

1. 신규 고객 유치 한계
    - 시장 포화로 인한 신규 고객 유치 어려움
    - 보험 가입 검진 시 높은 ‘거절’ 비율(18.9%)
2. 잠재 위험 고객 예측 실패로 인한 보험료 과다 지급
3. 보험사의 혈액 검사 비용 부담

진행 방향

- 혈액 검사를 최소화 하여 추가비용을 절약한다.
  - 고객의 기본 정보를 기반으로 질병을 예측한다. **(기본 질병 예측모델)**
  - 예측된 질병의 청구비 기대값을 기준으로 혈액 검사 시행 여부를 결정한다.
- 위험 고객 예측 모델을 향상시켜 보험료 지급을 최소화 한다.
  - 혈액 검사와 기본 정보를 기반으로 질병을 예측한다. **(혈액검사 질병 예측 모델)**
  - 질병 발생이 예측된 고객은 해당 질병에 따라 가입 여부를 처리한다.
  - 질병의 청구비 기대값이 높은 경우 승인, 낮은 경우 거절한다.

진행 방향


기본 정보
[키, 나이, 신장, 혈압, 맥박]
고객 청구비 기대값 책정
예측 질병 코드
[A, B, ..]
+
질병코드
보험비 평균
A
193,537
B
318,071
...
....
청구비 위험
청구비 낮음
청구비 적정
거절
승인
재검
혈액 검사 정보
청구비 위험
청구비 낮음
거절
승인

데이터 정제

데이터 수집

- **고객 검진 정보** (insu_pre_review)
- **보험료 청구,지급 정보** (insu_request)
- 보험 상품 가격 정보 (insu_price)
- 질병 코드 정보 (insu_code_sick)
- 국민 건강 검진 결과 (insu_nh_h_screen)

데이터 정제

데이터 수집

고객 검진 정보 (insu_pre_review) : 14,939개

변수	설명	속성
고객ID	일련번호	문자
검사구분	혈액검사 / 일반검사	문자
판정결과	승인, 거절, 재검	문자
검사일자	검사일자	날짜(숫자)
성별	남자 1, 여자 2	범주(숫자)
연령	연령	숫자
신장	신장	숫자
체중	체중	숫자
가슴둘레	가슴둘레	숫자
허리둘레	허리둘레	숫자

변수	설명	속성
혈압(이완기)	정상 : 80mmHg 미만	숫자
혈압(수축기)	정상 : 120mmHg 미만	숫자
맥박	정상 : 60-100회	숫자
총콜레스테롤	정상 : 30~135mg/dL	숫자
혈색소	정상 : 14.0~18.0 g	숫자
적혈구 혈색소	정상 : 4.2~5.4 e12/L	숫자
혈청 크레아티닌	정상 : 0.8~1.7mg/dL	숫자
B 형 감염 항원	음성 / 양성	문자
적혈구용적	정상 : 81 - 96 fl	숫자
적혈구 혈색소 농도	정상 : 8 g/dL	숫자

변수	설명	속성
PLAT.혈소판수	정상 : 4.2~5.4 e12/L	숫자
RBC.적혈구수	정상 : 4.2 ~ 5.4 x 1012/L /	숫자
WBC.백혈구수	정상 : 4.000 - 10,000/mm3	숫자
RGPT.감마 GPT	정상 : 0 - 40IU/L,	숫자
SGOT~AST	정상 : 0~40 IU/L	숫자
SGPT~ALT	정상 : 0~40 IU/L /	숫자
TRIG.중성지방	정상 : 150mg/dL 미만	숫자
식전혈당(공복혈당)	정상 : 100~125mg/dl	숫자
전체 판단 점수	혈액 검사 전체 결과 (0~10)	숫자

데이터 정제

데이터 수집

보험료 청구, 지급 정보(insu_request) : 49,450 개

변수	설명	속성
고객ID	각 고객의 고유 ID	문자
검사구분(가입 사전검사)	혈액검사 / 일반검진 으로 분류	문자
판정결과	재검 / 승인으로 분류	문자
성별	남자 1, 여자 2	범주(숫자)
연령(보험 가입시점)	가입 당시의 연령(5단위 X)	숫자
계약보험ID	가입자가 계약한 보험에 대한 ID	문자
보험가입진단일	보험을 가입한 날	날짜(숫자)
보험상품ID	고객이 가입한 보험 상품의 ID (상품 고유번호)	문자
보험상품명	고객이 가입한 상품	문자
청구번호	가입자가 청구하는 경우 해당 청구 건수를 묶은 고유 번호	문자
청구서순번	청구 번호 안에 존재하는 청구 내역(입원, 약, 수술 등)	숫자
주상병	가입자가 진료 기간 중 진단이나 치료 횟수가 빈번한 질환	문자

변수	설명	속성
상병코드1	중요도 1위	문자
상병코드2	중요도 2위	문자
상병코드3	중요도 3위	문자
진단유형	외래 (24시간 미만) / 입원 (24시간 이상 입원)	문자
진단시작일	day-month-year	날짜(숫자)
진단종료일	day-month-year	날짜(숫자)
진단기간	(진단 종료일 - 시작일) + 1	숫자
보험청구금액	청구액이 없는 경우는 정액 담보(예: 입원일당)로 지급한 경우	숫자
보험지급금액	청구번호별 총 지급금액	숫자
보험금지급일자	day-month-year	날짜(숫자)
보험상품 가입기간	단위 : 월	숫자
누적 납입 보험료	보험상품 가입기간 x 기본 보험료	숫자

데이터 정제

데이터 수집

보험 상품 가격 정보(insu_price)

변수	설명	속성
보험상품 ID	보험상품 ID는 보험 상품명에 1:1 대응이 됨	문자
보험상품명	보험상품 ID는 보험 상품명에 1:1 대응이 됨	문자
기본 보험료	계약 체결 시 매월 계속 납입하기로 한 월 보험료	숫자

데이터 정제

데이터 수집

질병 코드 정보 (insu_code_sick)

변수	설명	속성
상병코드(3)	상병명(한국어)	문자
상병코드(4)	상병명(영어)	문자
상병명(한국어)	적용성별	문자
상병명(영어)	ex) Cholera	문자
적용성별	1 : 남성 / 2: 여성 / NaN (결측치)	범주(숫자)
적용나이(상한)	15 20 55 24 5 NaN	숫자
적용나이(하한)	15 10 8 40 20 NaN	숫자

데이터 정제

고려사항

보험비 청구 정보

- 한 고객이 여러번의 청구를한다.
- 하나의 청구를 여러차례로 나눠서 청구하는 경우가 존재한다. (동일한 청구번호에 대해)
- 동일한 질병, 동일한 보험이여도 지급액이 다른 경우가 존재한다.

데이터 정제

결측치 처리

고객 검진 정보

- 일반검진을 한 고객의 경우 혈액검사 데이터에 결측이 있다.
  - **혈액 검사 고객 데이터**와 **일반 고객 데이터**로 분할하여 개별적으로 처리한다.

- 가슴둘레와 허리둘레에 결측된 데이터가 있다.(4)
  - 전체 데이터 수(14,939) 에 비해 소수 이므로 평균값으로 대체한다.

데이터 정제

결측치 처리

보험비 청구 정보

- 지불 금액(req_amount) 에 결측이 있다.(2)
  - 전체 데이터 수(49,450)에 비해 소수 이므로 결측 데이터를 제거한다.

- 질병코드가 ‘ZZZ’ 인 데이터는 결측치로 입력된 데이터이다.(7,185)
  - 질병을 기준으로 분석하기 때문에 해당 데이터는 사용할 수 없다.

- 2,3 상병에 결측이 있다.(11,499)
  - 2,3 상병 항목는 부가적(optional) 인 특성을 갖는다. 고객을 기준으로 1,2,3 상병을 묶어서 하나의 변수를 생성한다.

customer_id	sicks
C112379	S02/S92
C134227	R60/Z03/J00/J06/N28
C134251	R51

탐색적 분석

질병 코드 분석

검진 데이터로 예측할 불가능한 질병 제외.

- F, H, L, M, P, Q, R, S, T, U, V, Y, Z

코드분류	설명
A	특정 감염성 및 기생충성 질환
B	특정 감염성 및 기생충성 질환
C	신생물(C00-D48)
D	혈액 및 조혈기관의 질환과 면역메커니즘을 침범한 특정 장애(D50-D89)
E	내분비, 영양 및 대사 질환
F	정신 및 행동 장애
G	신경계통의 질환
H	눈 및 눈 부속기 / 귀 및 유돌의 질환
I	순환계통의 질환
J	호흡계통의 질환
K	소화계통의 질환

코드분류	설명
L	피부 및 피하조직의 질환
M	근골격계통 및 결합조직의 질환
N	비뇨생식계통의 질환
O	임신, 출산 및 산후기
P	출생전후기에 기원한 특정 병태
Q	선천기형, 변형 및 염색체이상
R	달리 분류되지 않은 증상, 징후와 임상 및 검사의 이상소견
S,T	손상, 중독 및 외인에 의한 특정 기타 결과
U	특수목적 코드
V,Y	질병이환 및 사망의 외인
Z	건강상태 및 보건서비스 접촉에 영향을 주는 요인

탐색적 분석

질병 코드 분석

혈액 검사로 예측가능한 질병 코드를 선별.

- A, B, C, D, E, G, I, J, N (출처 : 국민건강지식센터)

코드분류	설명
A	특정 감염성 및 기생충성 질환
B	특정 감염성 및 기생충성 질환
C	신생물(C00-D48)
D	혈액 및 조혈기관의 질환과 면역메커니즘을 침범한 특정 장애(D50-D89)
E	내분비, 영양 및 대사 질환
F	정신 및 행동 장애
G	신경계통의 질환
H	눈 및 눈 부속기 / 귀 및 유돌의 질환
I	순환계통의 질환
J	호흡계통의 질환
K	소화계통의 질환

코드분류	설명
L	피부 및 피하조직의 질환
M	근골격계통 및 결합조직의 질환
N	비뇨생식계통의 질환
O	임신, 출산 및 산후기
P	출생전후기에 기원한 특정 병태
Q	선천기형, 변형 및 염색체이상
R	달리 분류되지 않은 증상, 징후와 임상 및 검사의 이상소견
S,T	손상, 중독 및 외인에 의한 특정 기타 결과
U	특수목적 코드
V,Y	질병이환 및 사망의 외인
Z	건강상태 및 보건서비스 접촉에 영향을 주는 요인

탐색적 분석

청구,지불 데이터 분석

- **보험비 청구 정보**에서 빈도 높은 질병 코드를 분석하였다.

질병코드	설명	청구 수
S33	요추 및 골반의 관절 및 인대의 탈구, 염좌 및 긴장	504
M79	달리 분류되지 않은 기타 연조직장애	503
J00	급성 비인두염[감기]	447
R10	복부 및 골반 통증	434
K29	위염 및 십이지장염	313
M75	어깨병변	313
M54	등통증	295
N64	유방의 기타 장애	283
Z03	의심되는 질병 및 병태를 위한 의학적 관찰 및 평가	274
S13	목부위의 관절 및 인대의 탈구, 염좌 및 긴장	265
R51	두통	247
K30	기능성 소화불량	235
E07	갑상선의 기타 장애	228
N76	질 및 외음부의 기타 염증	223

혈액검사로 판단할 수 없는 질병이 다수 존재.

- 해당 데이터를 질병 예측 모델 학습에 사용하는 경우, 모델의 정확도를 저하시킨다.

- 혈액 검사로 예측 가능한 청구 데이터만 고려한다.

탐색적 분석

청구,지불 데이터 분석

- **보험비 청구 정보**에서 청구 금액이 높은 질병 코드를 분석하였다.

질병코드	청구 금액(평균)	청구 수
C85	8692878.0	1
D05	5169884.0	1
I64	4021562.0	3
J86	3640378.0	1
Z95	3430883.0	1
C96	3361508.0	1
C61	3128496.0	1
Q23	3112782.0	1
M87	3099960.0	1
S12	3099655.0	1
S51	2988018.0	1
G00	2889380.0	1
I61	2758872.0	4
K83	2713979.0	2

청구 금액이 높은 질병들은 청구 빈도가 낮다.
- 청구 금액을 기준으로 분석하기에는
데이터가 적어 일반화가 어렵다.

탐색적 분석

청구,지불 데이터 분석

- **보험비 청구 정보**에서 질병 코드 별 청구 금액과 청구 빈도를 분석하였다.

질병코드	청구 금액 (평균)	청구 빈도
A	193,537	165
B	318,071	190
C	1,401,420	160
D	491,425	508
E	149,703	638
G	756,203	203
I	752,848	504
J	352,853	1180
N	202,222	1423

각 질병 코드의 빈도가 어느정도 있기 때문에

질병 별로 모델을 학습할 수 있다.

탐색적 분석

청구,지불 데이터 분석

- 질병 코드 데이터 수를 분석하였다.

질병코드	청구 수
A	199
B	181
C	121
D	486
E	503
F	8
G	211
H	411
I	418

질병코드	청구 수
J	835
K	1,312
L	246
M	1,600
N	961
O	3
P	1
Q	7
R	1,390

질병코드	청구 수
S	1,773
T	351
U	6
V	3
W	10
X	11
Y	1
Z	343

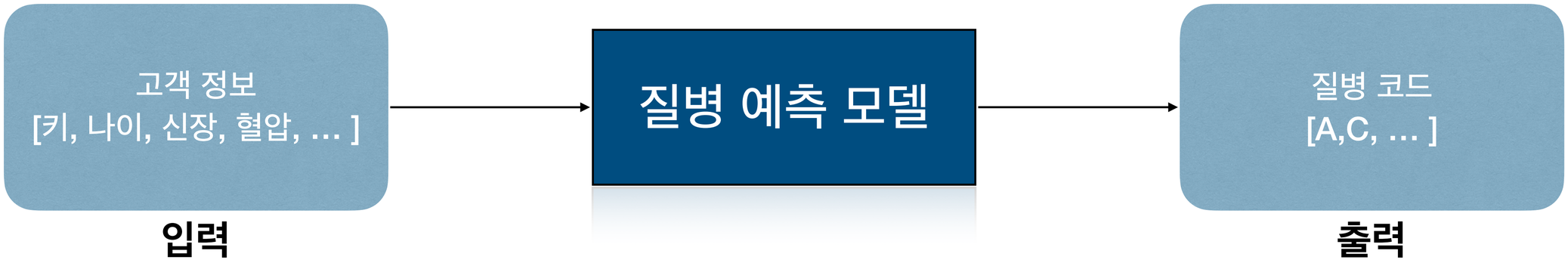
질병코드 F, O, P, Q, U, V, W, X, Y 의 경우

데이터 수가 적어 분석하기에 어려움이 있다.

모델링

계획

- **고객 건강 정보**로 가능성 있는 **질병들을 예측**하는 모델을 학습시킨다.



모델링

질병 예측 모델

- **질병 대분류 코드** (알파벳)가 목표 변수 이다.
- **기본(일반 고객+혈액검사 고객)고객과 혈액검사 고객** 별로 모델을 생성한다.
- 기본 모델에서는 **성별, 나이, 몸무게, 가슴, 허리 둘레**가 설명 변수이다.
- 기본 모델에서는 질병코드 **A, B, C, D, E, G, I, J, K, N** 를 목표 변수로 한다.
- 혈액 모델에서는 **혈액 검사 점수**(19개의 항목) 이 설명 변수이다.
- 혈액 모델에서는 질병코드 **A, B, C, D, E, G, I, J, N** 를 목표 변수로 한다.
- 각 질병 코드 별로 예측 결과를 출력한다.
- 질병마다 Decision Tree 와 Random Forest, Gradient Boosting 중 높은 점수의 모델을 선정한다.

모델링

기본 모델

- 데이터 수 : 14,932
- 목표변수 : A, B, C, D, E, G, I, J, K, N
- 설명변수 : 성별, 나이, 몸무게, 가슴, 허리 둘레, 혈압, 맥박
- 데이터 분할 : Train(0.6) / Validation(0.2) / Test(0.2)
- 평가지표 : Precision - 고객이 병에 걸릴 경우 risk가 크므로 Precision 을 사용한다.

모델링

기본 모델

- **Decision Tree**

질병코드	Train	Test	Precision
A	1.0	0.965	0.697
B	1.0	0.967	0.709
C	1.0	0.973	0.634
D	1.0	0.92	0.722
E	1.0	0.916	0.720
G	1.0	0.962	0.699
I	1.0	0.932	0.730
J	1.0	0.871	0.746
K	1.0	0.831	0.709
N	1.0	0.866	0.767

- 비교적 성능이 낮지만, 전체적으로 높은 성능을 보인다.

모델링

기본 모델

- **Random Forest**

질병코드	Train	Test	Precision
A	1.0	0.987	1.0
B	1.0	0.985	1.0
C	1.0	0.99	1.0
D	1.0	0.97	1.0
E	1.0	0.968	1.0
G	1.0	0.986	1.0
I	1.0	0.972	1.0
J	1.0	0.945	1.0
K	1.0	0.917	0.999
N	1.0	0.934	0.934

- DT 에 비해 성능이 높다.

모델링

기본 모델

- **Gradient Boosting**

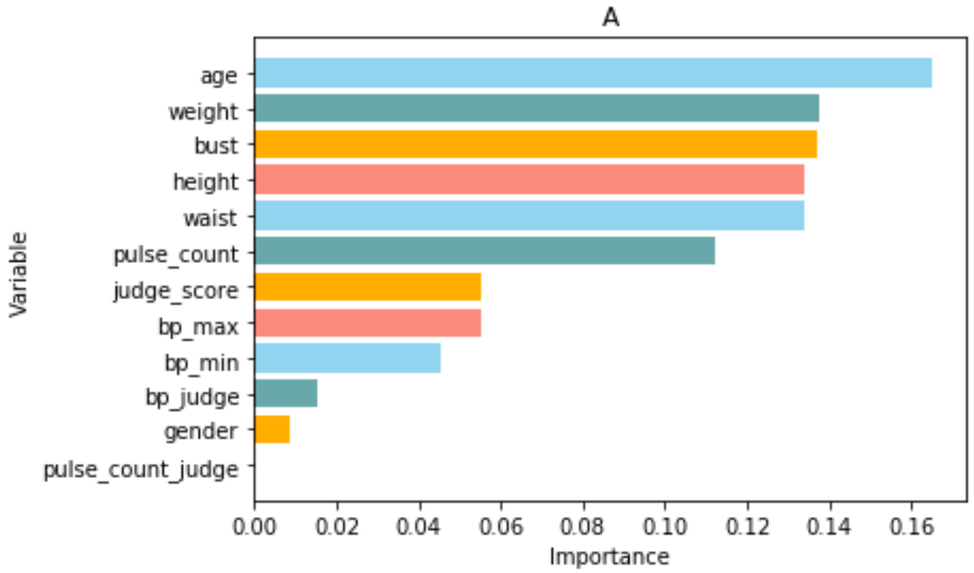
질병코드	Train	Test	Precision
A	0.988	0.988	1.0
B	0.989	0.984	0.75
C	0.993	0.99	1.0
D	0.968	0.969	0.818
E	0.967	0.968	0.888
G	0.986	0.986	0.857
I	0.973	0.971	0.9
J	0.945	0.944	0.888
K	0.912	0.917	1.0
N	0.937	0.934	1.0

- 전체적인 Score 는 Random Forest 와 유사하다.

모델링

기본 모델

- **설명 변수 중요도**



- 모델 별, 질병코드 별로 변수 중요도에 약간의 차이가 보인다.
- 일반적으로 나이, 몸무게, 신장이 높은 중요도를 보인다.

모델링

혈액 모델

- 데이터 수 : 3,191
- 목표변수 : A, B, C, D, E, G, I, J, N 질병유무(0/1)
- 설명변수 : 혈액검사 점수(19개 - 0/1)
- 데이터 분할 : Train(50%) / Validation(20%) / Test(30%)
- 평가지표 : Precision - 고객이 병에 걸릴 경우 risk가 크므로 Precision 을 사용한다.

모델링

혈액 모델

- **Decision Tree**

질병코드	Train	Test	Precision
A	0.957	0.931	0.538
B	0.953	0.973	0.857
C	0.969	0.964	1.0
D	0.889	0.851	0.611
E	0.869	0.862	0.655
G	0.952	0.934	0.6
I	0.899	0.892	0.64
J	0.806	0.781	0.771
N	0.773	0.757	0.75

- 비교적 성능이 낮지만, 전체적으로 높은 성능을 보인다.

모델링

혈액 모델

- **Random Forest**

질병코드	Train	Test	Precision
A	0.957	0.937	0.777
B	0.953	0.975	1.0
C	0.969	0.964	1.0
D	0.889	0.857	0.666
E	0.869	0.873	0.863
G	0.952	0.936	0.75
I	0.899	0.901	0.809
J	0.806	0.787	0.744
N	0.773	0.757	0.714

- A, B, C, G, I, J 질병에 대해 점수가 높다.

모델링

혈액 모델

- **Gradient Boosting**

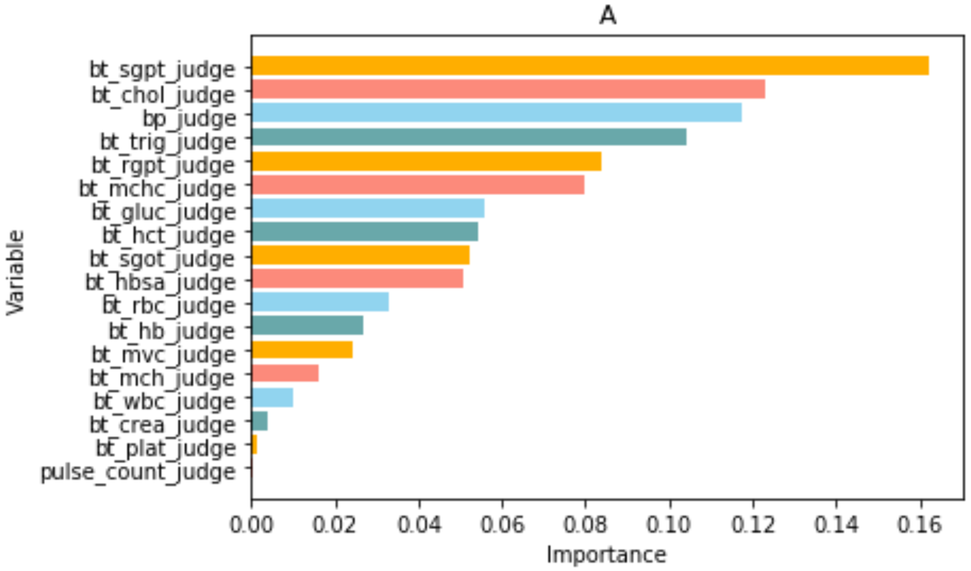
질병코드	Train	Test	Precision
A	0.955	0.937	0.5
B	0.953	0.975	1.0
C	0.969	0.961	0.333
D	0.886	0.859	0.666
E	0.864	0.875	0.714
G	0.952	0.936	0.666
I	0.895	0.901	0.571
J	0.801	0.787	0.761
N	0.767	0.76	0.692

- 일부 질병에서는 Random Forest 보다 높은 점수를 보인다.

모델링

혈액 모델

- **설명 변수 중요도**



- 질병코드 별 변수 중요도에 차이가 있다.

모델링

학습 모델 정리

- **기본 모델**

질병코드	모델
A	Random Forest
B	Random Forest
C	Random Forest
D	Random Forest
E	Random Forest
G	Random Forest
I	Random Forest
J	Random Forest
K	Gradient Boosting
N	Gradient Boosting

- **혈액 모델**

질병코드	모델
A	Random Forest
B	Random Forest
C	Random Forest
D	Gradient Boosting
E	Random Forest
G	Random Forest
I	Random Forest
J	Gradient Boosting
N	Random Forest

결론

- 나이, 성별, 신장 등 기본적인 정보로 질병을 예측하는 **기본 모델** 과
  혈액검사 정보로 질병을 예측하는 **혈액 모델** 을 생성하였다.
- 각 질병군 마다 하나의 모델을 생성하였다.
- 질병, 모델 별 성능의 차이가 있었고, **Precision** 을 기준으로 사용할 모델을 선정하였다.
- 실제 적용시 모델의 **예측 정보**와 **질병군 청구 기대값** 요소를 고려하여 의사결정을 할 수 있다.