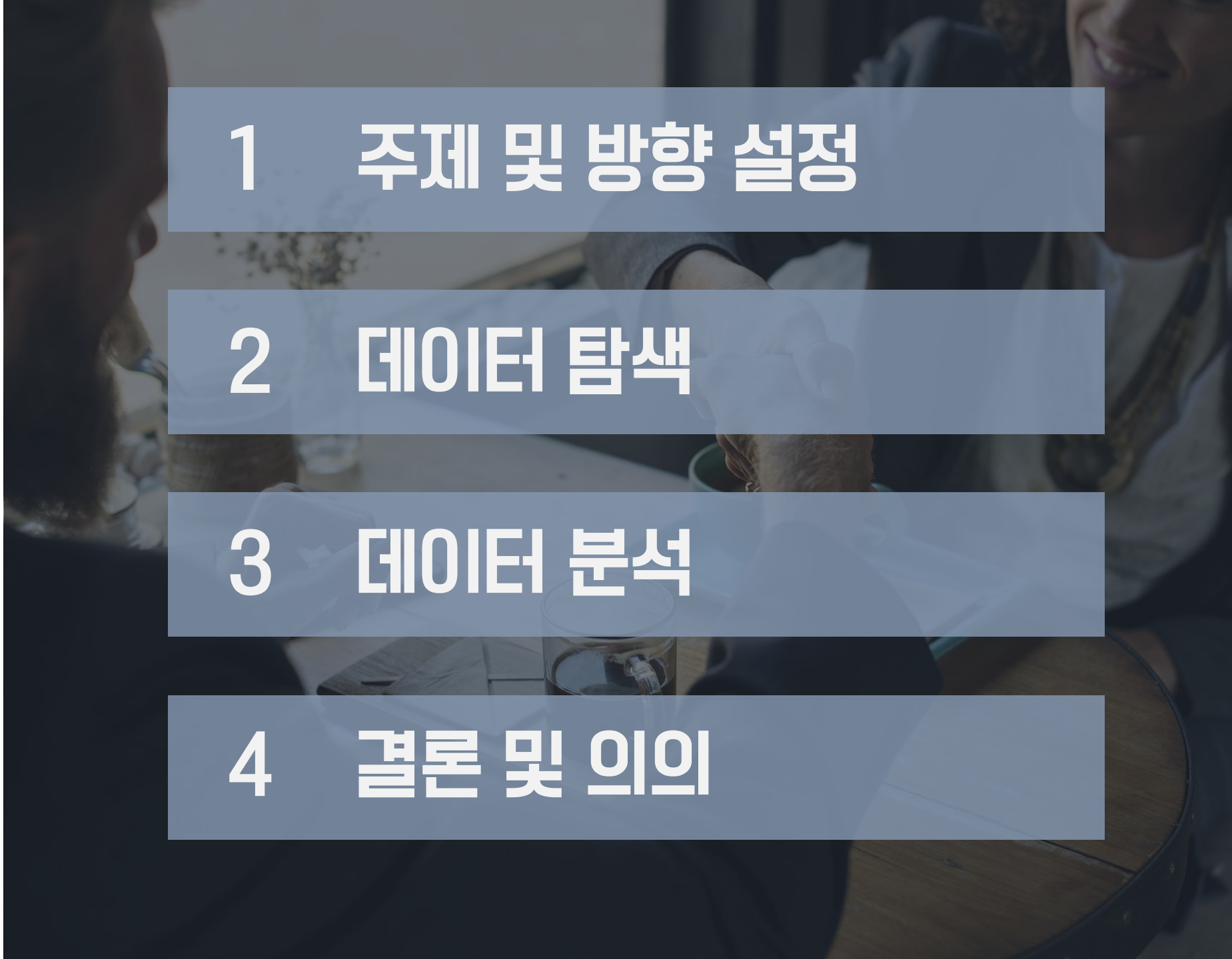
대출연체 예측 알고리즘 개발

KingoFive

일반화선형모델을 통한 해석과
앙상블 기법을 활용한 연체예측

Contents



1

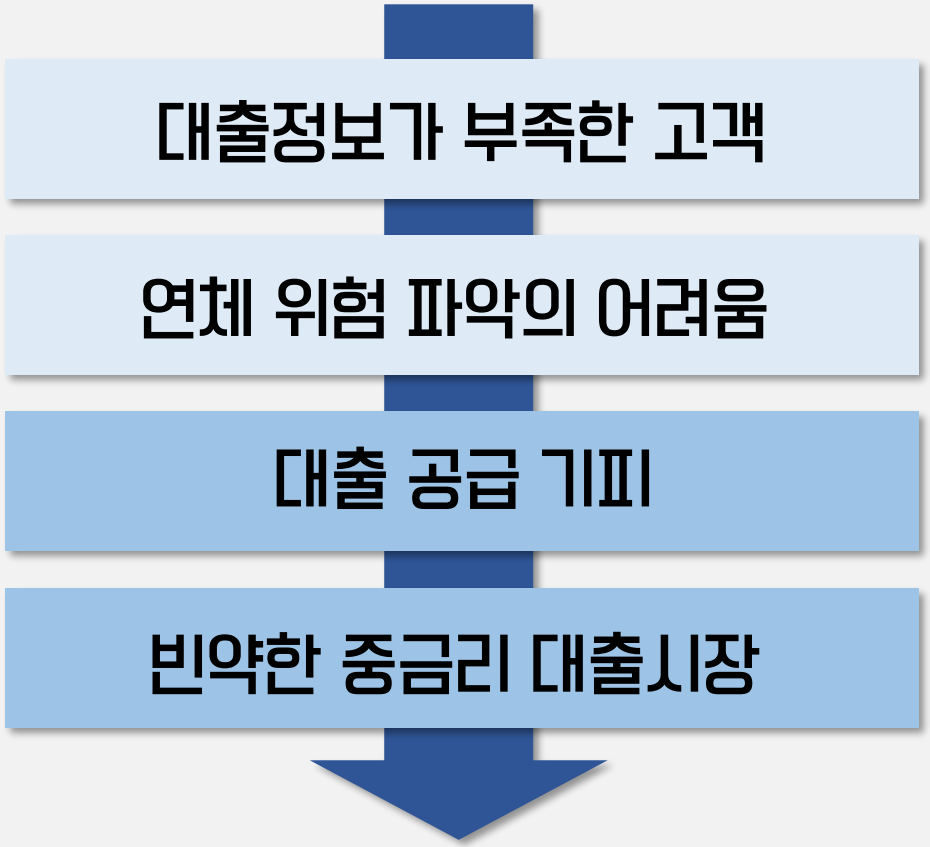
주제 및
방향 설정

문제상황의 인식 및 주제 설정

기존 신용평가 방법의 한계

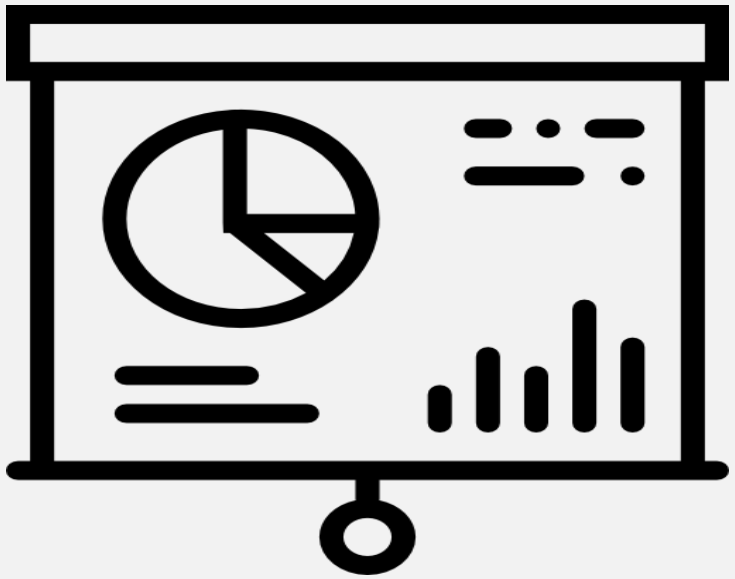
대출 정보가 존재하지 않는 사람의 상환 능력과 의지는
기존의 신용평가 방법으로는 효과적으로 평가할 수 없다.

문제상황의 인식 및 주제 설정



대출정보가 부족한 고객들의 상환 능력과 의지를 파악할 수 있는
변수들을 해석하고 고객의 연체 위험을 평가하는 방법을 제시

분석 방향



정교화된 고객 평가

주어진 정보로 연체 위험률 계산

Interpretation

해석모형을 통한 요인분석

→ 어떤 요인이 유의한가?
→ 요인의 정량적 위험은?

Prediction

예측 모형(머신 러닝)을 통한 예측

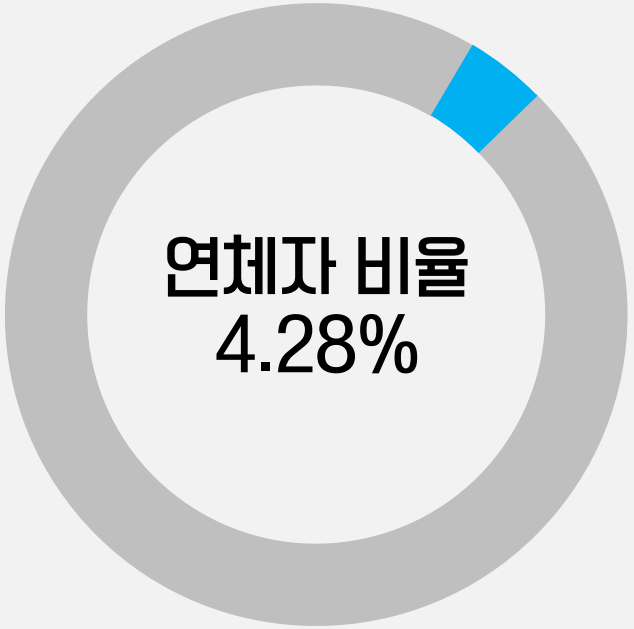
→ 모든 변수 사용(유의성이 낮은 요인의 정보량까지 포함)
→ 다양한 변수 조합 및 모형 고려

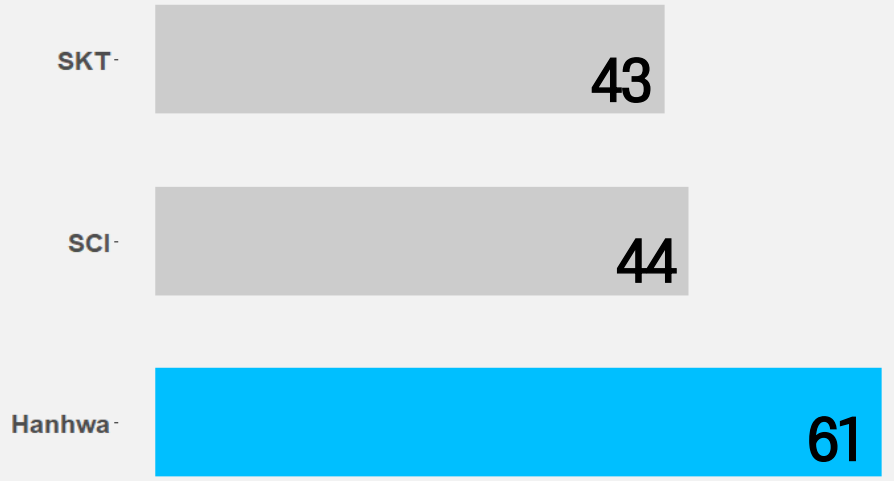
2

데이터 탐색

데이터 탐색

반응변수와 설명변수 내
0의 비중이 큼을 확인





데이터 제공기관 별 변수 내 평균 0비율(%)

- 총 100,233건의 데이터 중 연체자는 총 4,287건 뿐
  → 데이터의 비대칭성 확인

- 전체적으로 데이터에 0의 비중이 크고,
  그 중 한화 데이터에 0 비중이 큰 변수 많음

데이터 전처리 방향

1. 논리성

논리적으로 이치에 맞지 않는 부분들을 수정

2. 일관성

통일된 단위를 갖도록 전처리

3. 데이터 손실의 최소화

데이터 삭제보다는 유의미한 의미를 이끌어낼 수 있게 수정

데이터 전처리

총 대출 건수 & 총 대출 금액

- 총 대출건수가 0인 고객의 총 대출금액, 총 신용대출금액 모두 0으로 바꿔 줌

최대 월납입 보험료

- 월납 최대액이 00이면 보험료 연체율, 최근1년 보험료 연체율 0으로 수정

보험금 청구 & 보험지급액

- 보험금 청구건수가 0인 고객의 보험지급 금액을 0으로 처리

데이터 전처리

최초대출날짜 & 가입연월

- 기준일에서 값을 뺌으로써 기간으로 바꿔 줌

최근1년 보험료 연체율

- 숫자로 값을 바꾸고, 각 범주 구간의 중간값으로 바꿔 줌

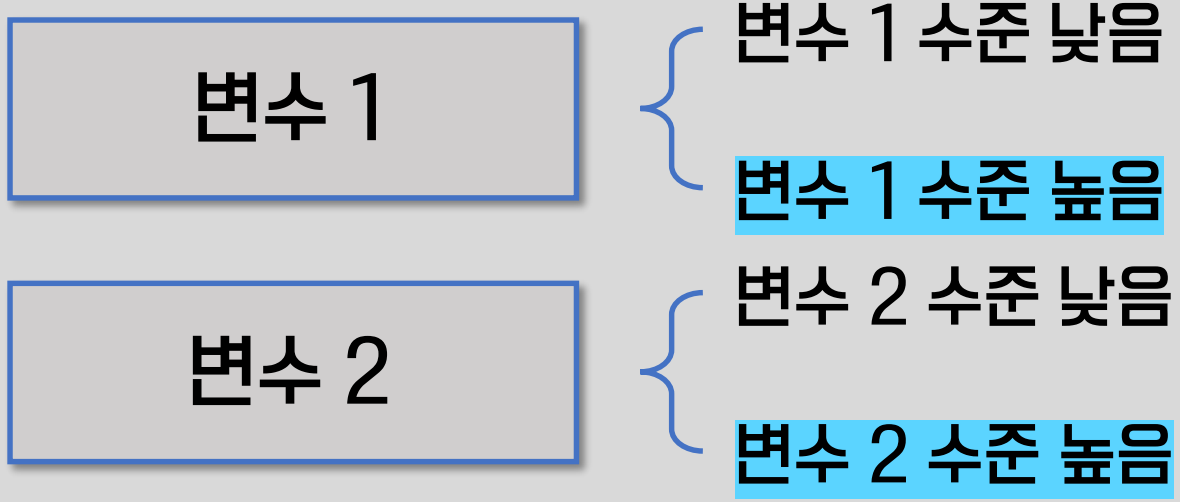
막내 자녀 나이

- 막내자녀나이 변수를 자녀의 유무(0,1)로 변형

전체 금액관련 변수

- '원'을 기본단위로 하여 금액 단위를 수정

파생변수 생성: 필요성



두 변수의 수준이 모두 높을 때 의미가 추가로 생성되는 변수들 존재
→ '두 개 모두에 포함된다'라는 정보를 담아 줄 변수 필요

파생변수 생성 Idea

1. 신용이 낮은 사람은 전체 대출액이 적고,
그 중에서도 은행에서의 대출액이 더 적을 것이다.

2. 2산업분류나 기타금융권에서의 대출 횟수가
1금융권보다 많다면 연체 가능성이 높을 것이다.

3. 기존에 통신비를 많이 연체하던 사람이 최근에도
많이 연체했으면 연체 가능성이 높을 것이다.

4. 보험료 완납 & 보험료 자동이체와 같은
정기 납부에 대한 실패가 많을 수록
연체 가능성이 높을 것이다.

파생변수 생성 Idea

1. 신용이 낮은 사람은 전체 대출액이 적고,
그 중에서도 은행에서의 대출액이 더 적을 것이다.

2. 2산업분류나 기타금융권에서의 대출 횟수가
1금융권보다 많다면 연체 가능성이 높을 것이다.

높고 낮음, 많고 적음을 어떻게 정해 분할할수 있을까?

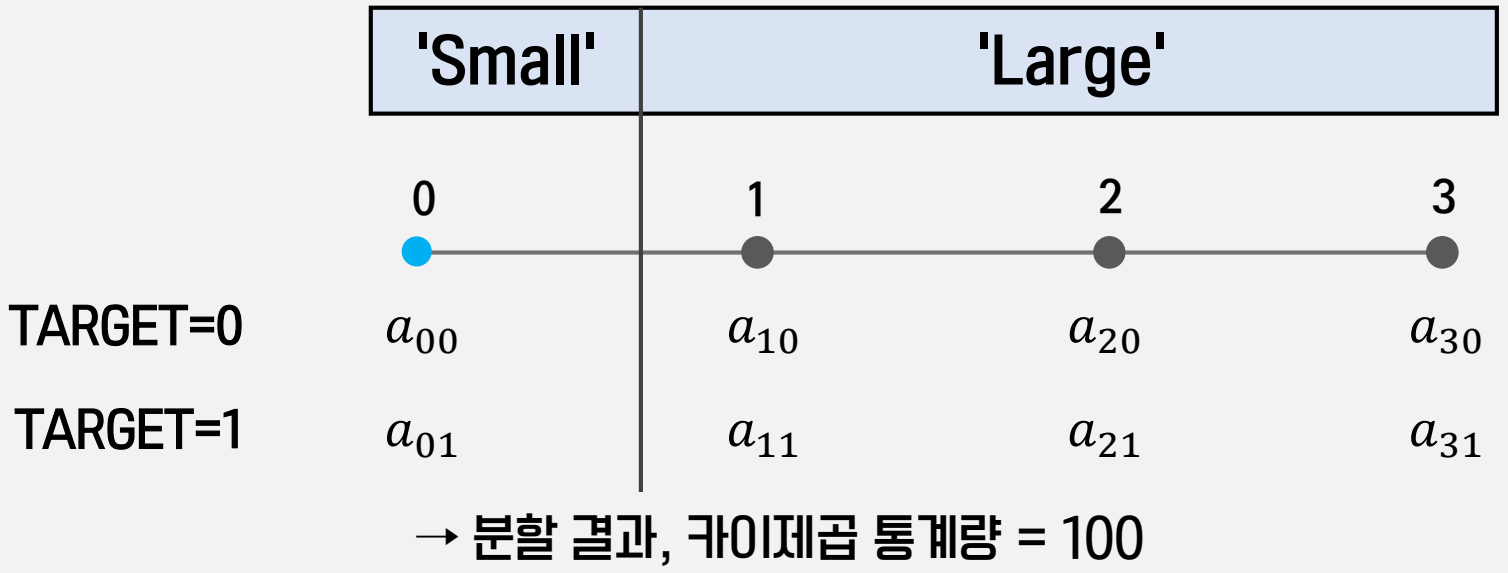
3. 기존에 통신비를 많이 연체하던 사람이 최근에도
많이 연체했으면 연체 가능성이 높을 것이다.

4. 보험료 완납 & 보험료 자동이체와 같은
정기 납부에 대한 실패가 많을 수록
연체 가능성이 높을 것이다.

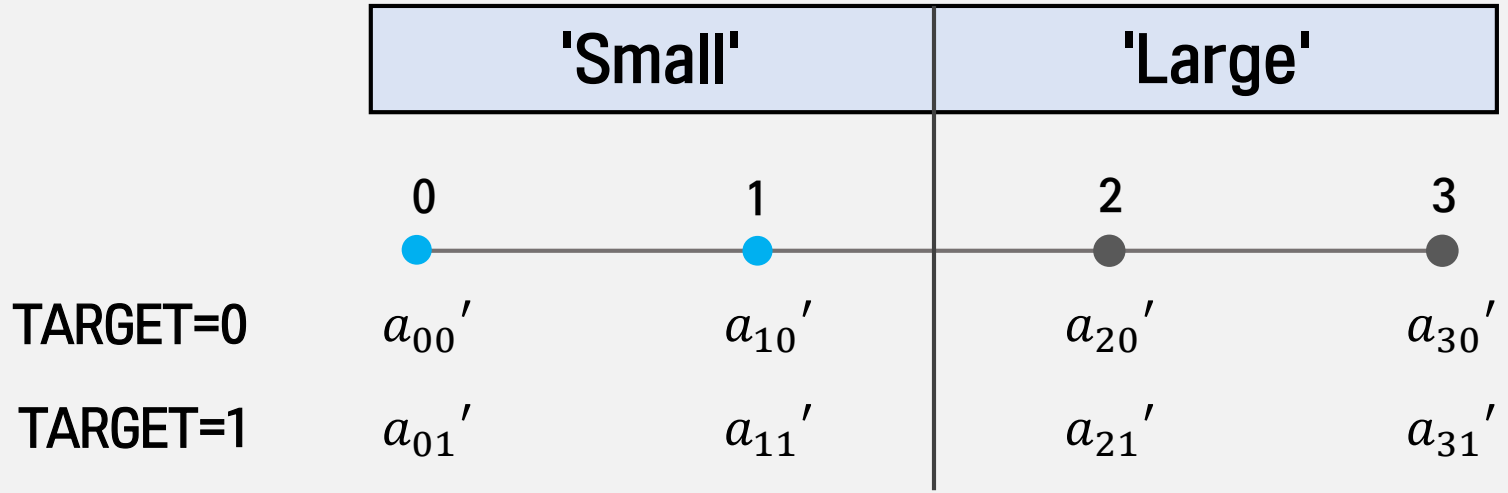
파생변수 생성:
카이제곱 분할

분할표 검정 개념을 활용한
최적의 분할지점 찾기

카이제곱통계량
분할의 결과가 연체여부와 상관이
얼마나 있는가를 나타내는 수치



→ 분할 결과, 카이제곱 통계량 = 100



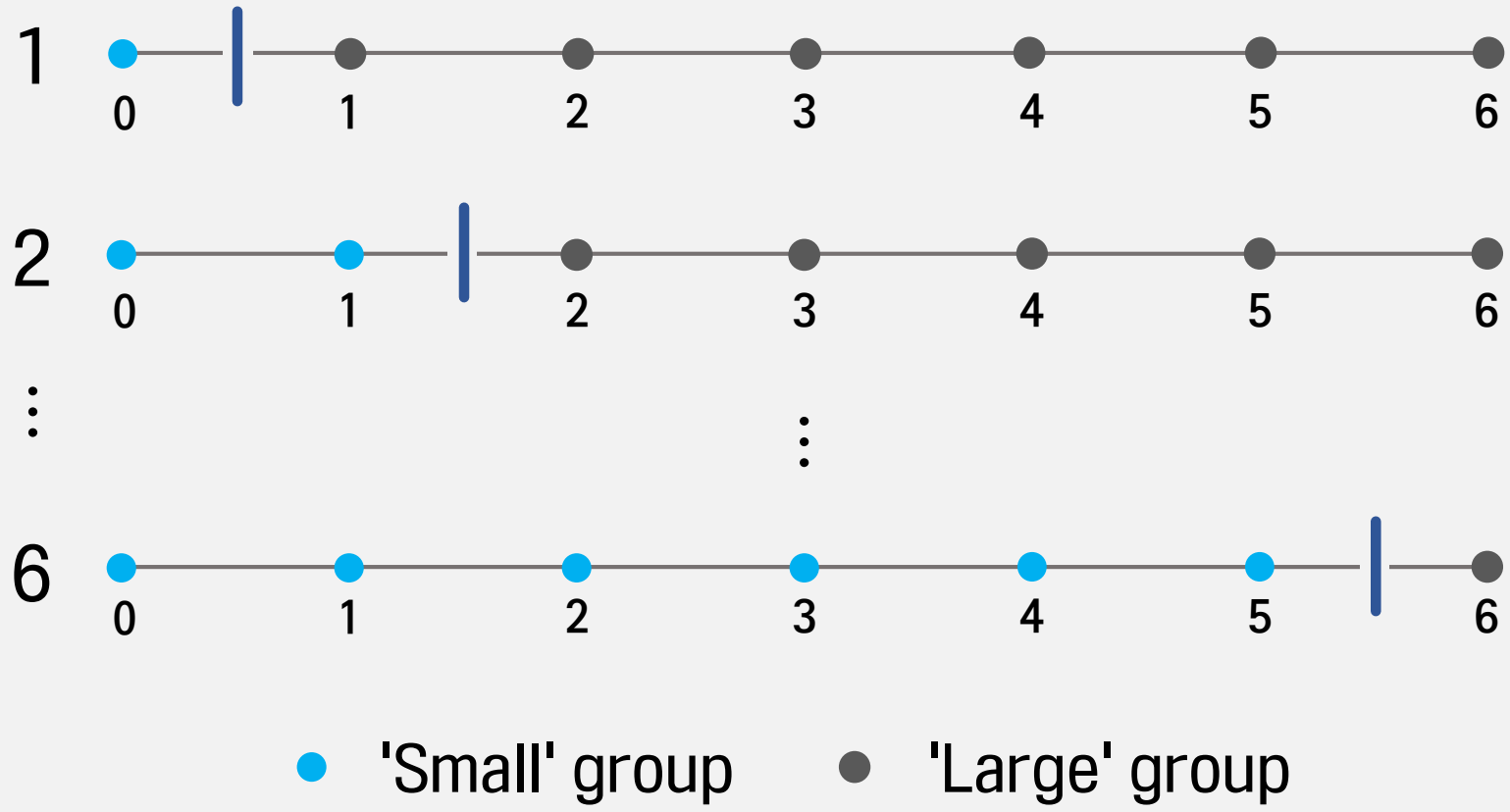
→ 분할 결과, 카이제곱 통계량 = 200

상관성의 정도가 더 큰 {0,1}, {2,3}으로 분할!

파생변수 생성:
카이제곱 분할

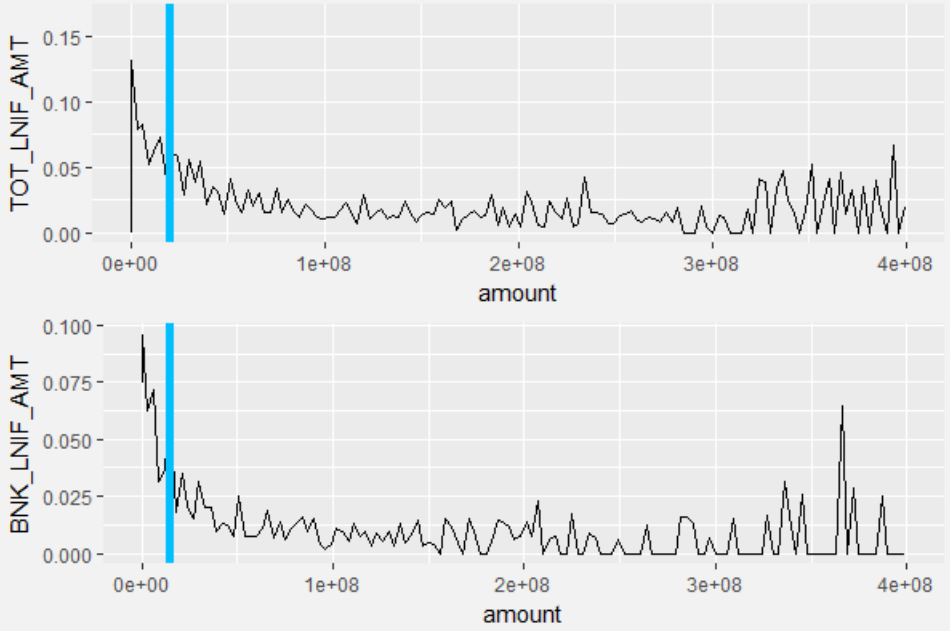
ex. 기타금융권 대출건수

카이제곱 통계량이 가장 큰 지점
(size 2, chi 687.09)을
최적의 분할로 선택

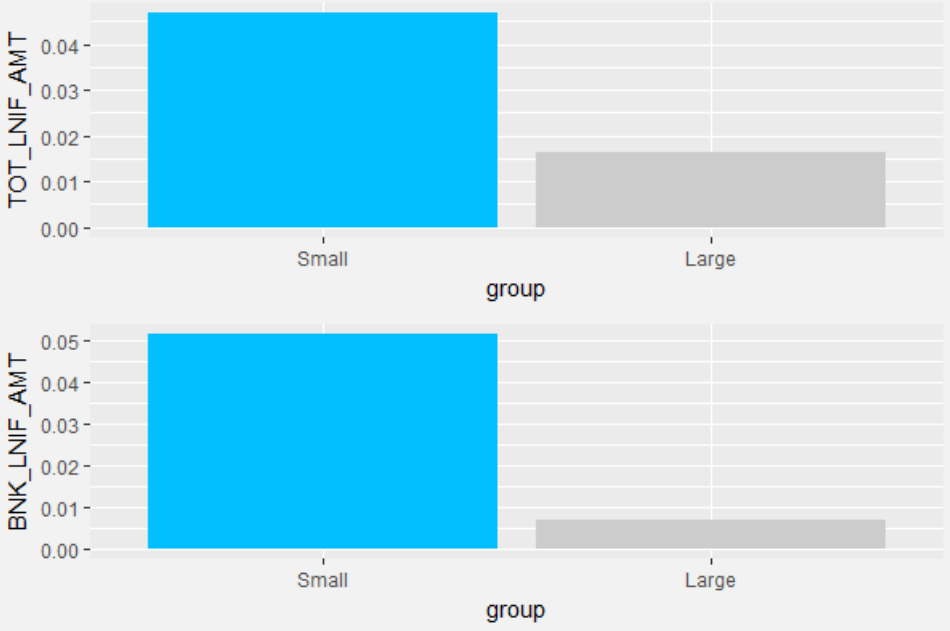


'Small' group size	1	2	3	4	5	6
Chi-square	502.32	687.09	292.77	77.23	0.84	1.41

파생변수 생성 방향



< 전체 대출액과 은행 대출액 금액 별 연체자 비중 >



< 대출액 적은 집단과 많은 집단 별 연체자 비중 >

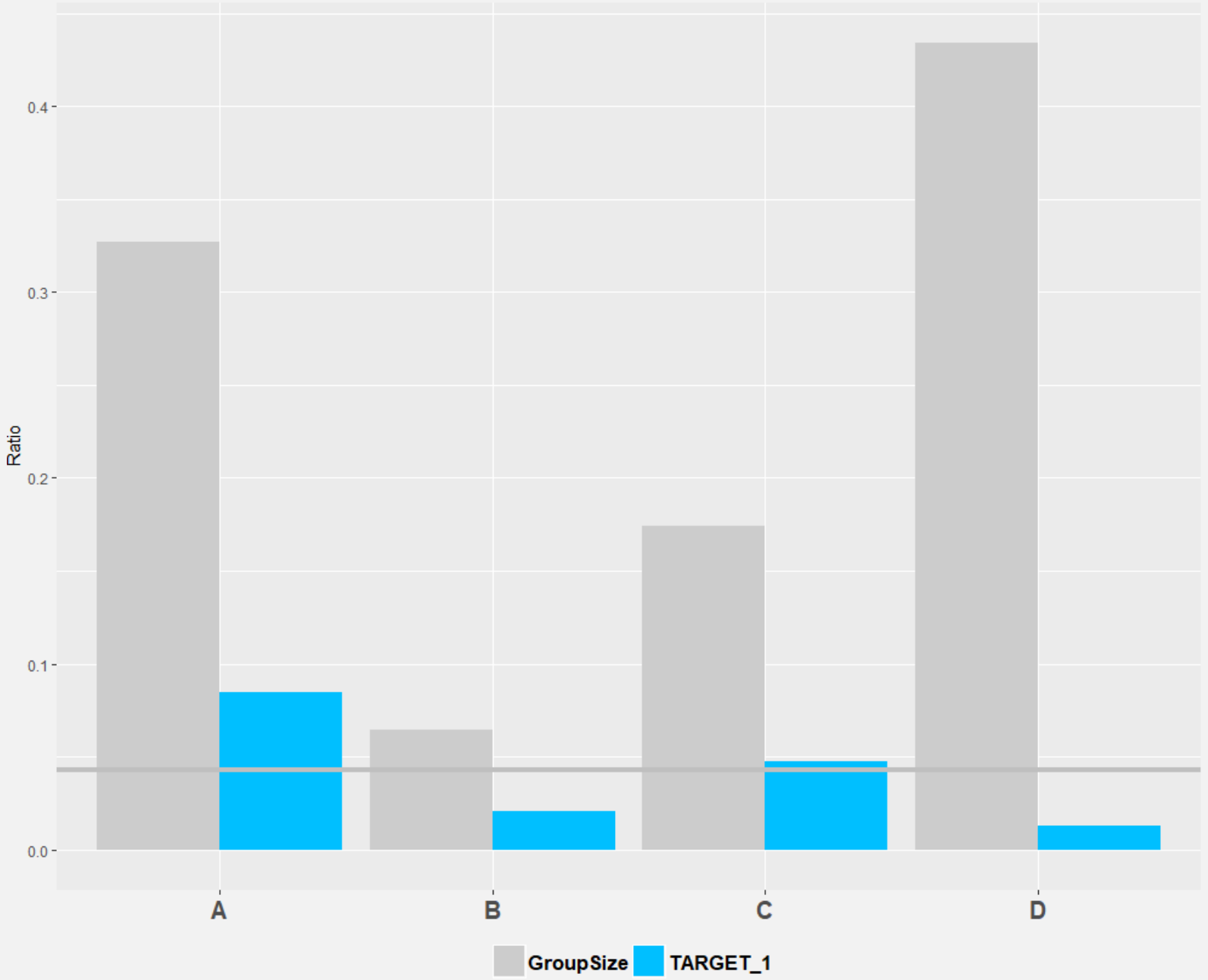
많음, 적음의 기준을 생성해 분할 → 분할한 변수끼리의 교호작용 탐색

분할을 통한 파생변수:
1.대출금액 비교

* 그래프의 회색 선은
  전체 데이터의 1비율인 0.043

총 대출 금액
X
은행 대출 금액

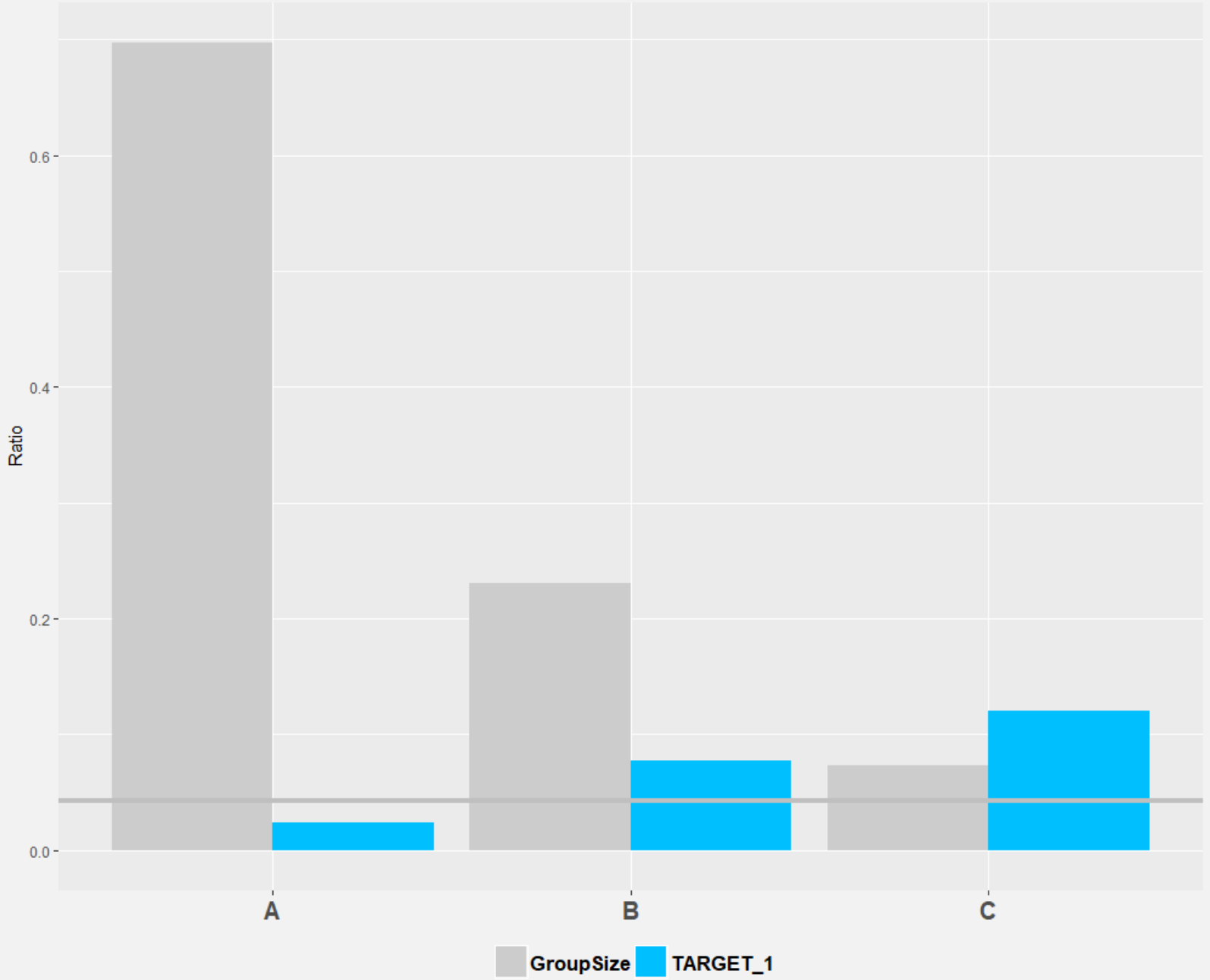
A : 전체, 은행 대출액 모두 적음
B : 전체 대출액 적고 은행 대출액 많음
C : 전체 대출액 많고 은행 대출액 적음
D : 전체, 은행 대출액 모두 많음



분할을 통한 파생변수:
2.대출기관 효과

은행 대출횟수
X
2산업분류 대출횟수
X
기타금융 대출횟수

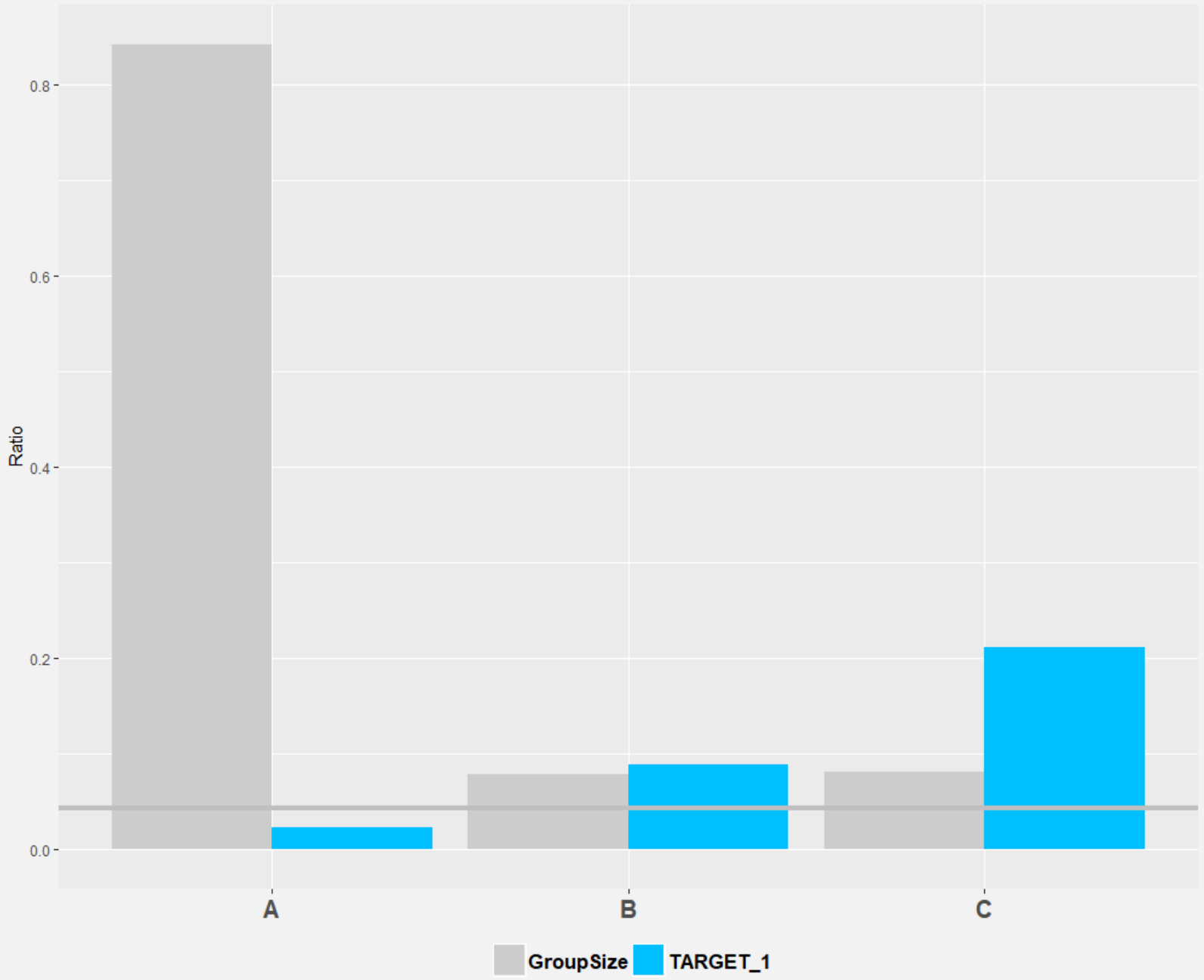
A : 은행 대출 건수 많은 그룹
B : 2산업분류 대출 건수 많은 그룹
C : 기타 금융권 대출 건수 많은 그룹



분할을 통한 파생변수:
3.통신비 연체 경향 효과

통신비 당월 연체 금액
X
통신비 연 최대 연체 금액

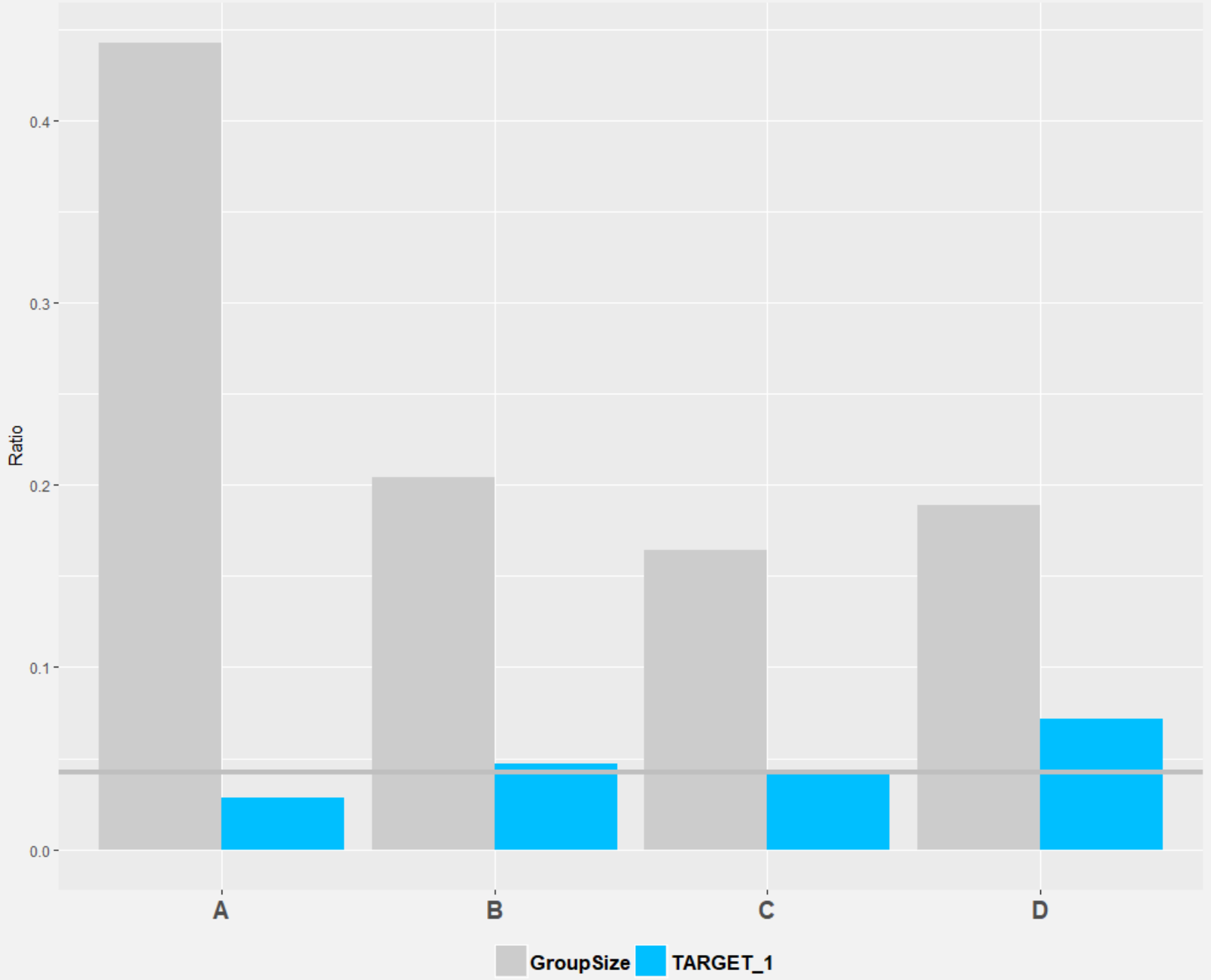
A : 연간 최대, 당월 연체액 모두 작음
B : 연간 최대 크고 당월 작거나
    연간 최대 작고 당월 큼
C : 연간 최대, 당월 연체액 모두 큼



분할을 통한 파생변수:
4.정기납부 연체효과

보험료 완납경험 횟수
X
보험료 자동이체 실패 월 수

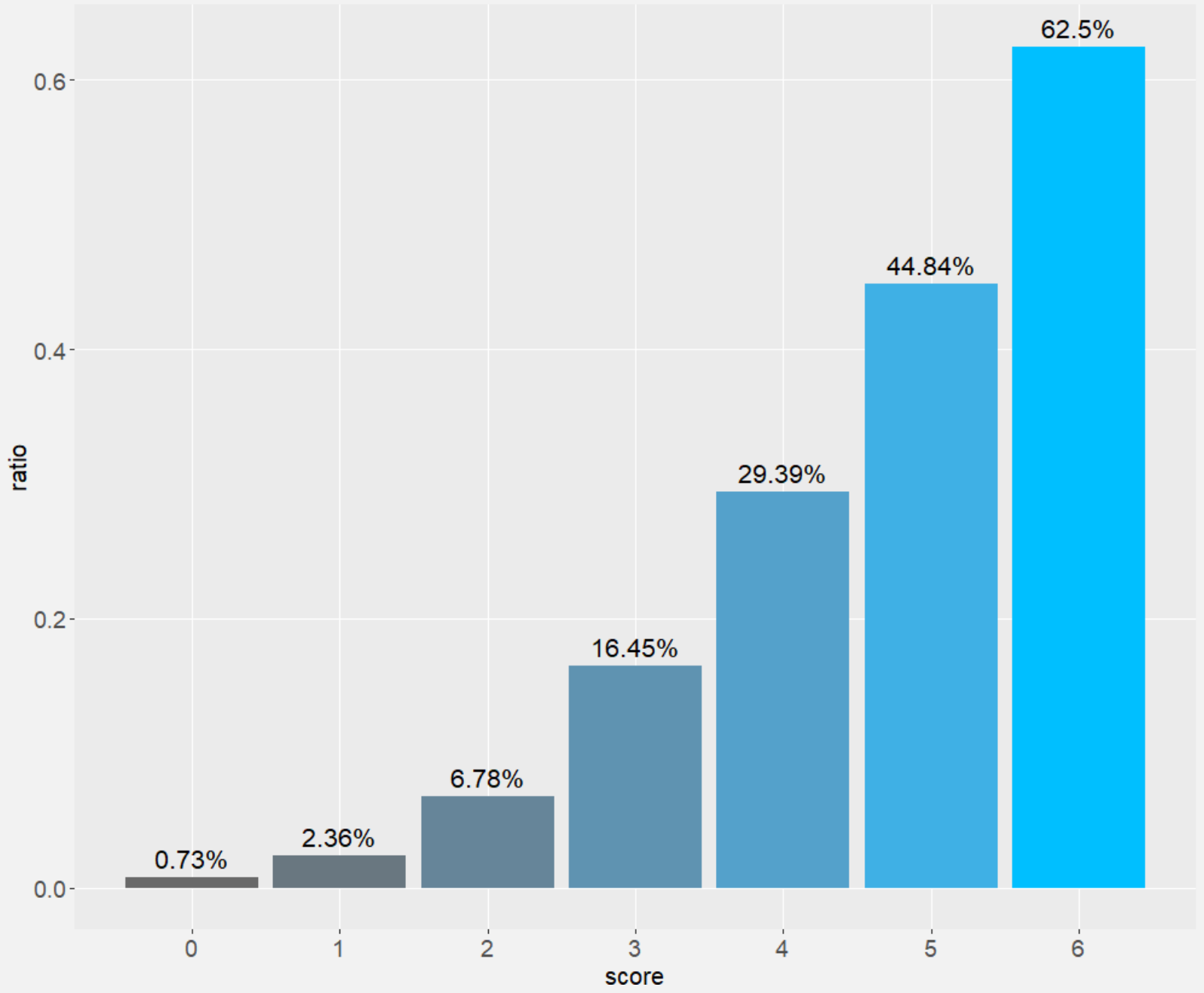
A : 이체실패경험 적고 완납실패 적음
B : 이체실패경험 많고 완납실패 적음
C : 이체실패경험 적고 완납실패 많음
D : 이체실패경험 많고 완납실패 많음



파생변수 생성 결과

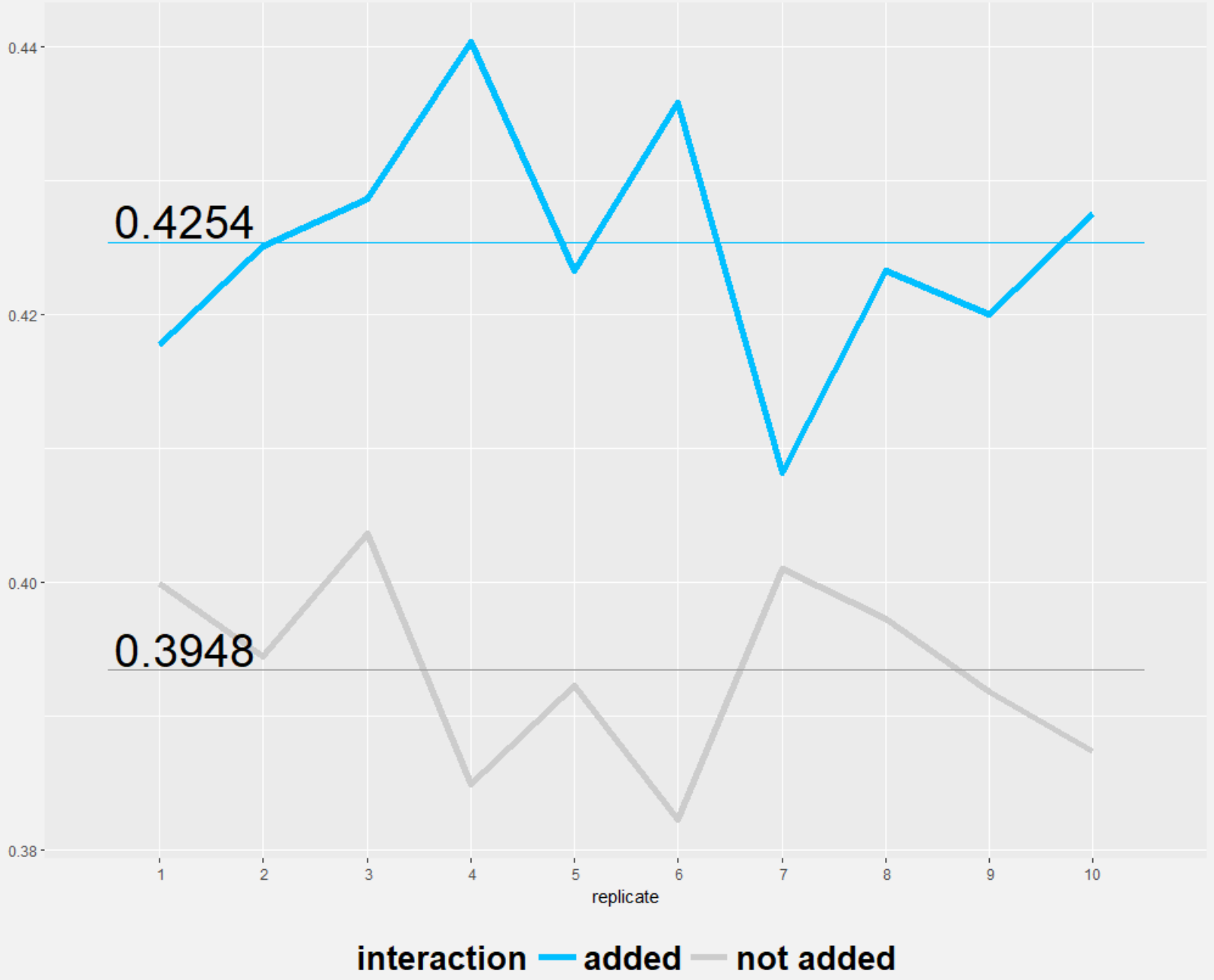
총 6개의 파생변수를 생성
→ 6개의 고위험 집단

0개부터 6개 집단 별 연체자 비중



파생변수 포함 여부 비교

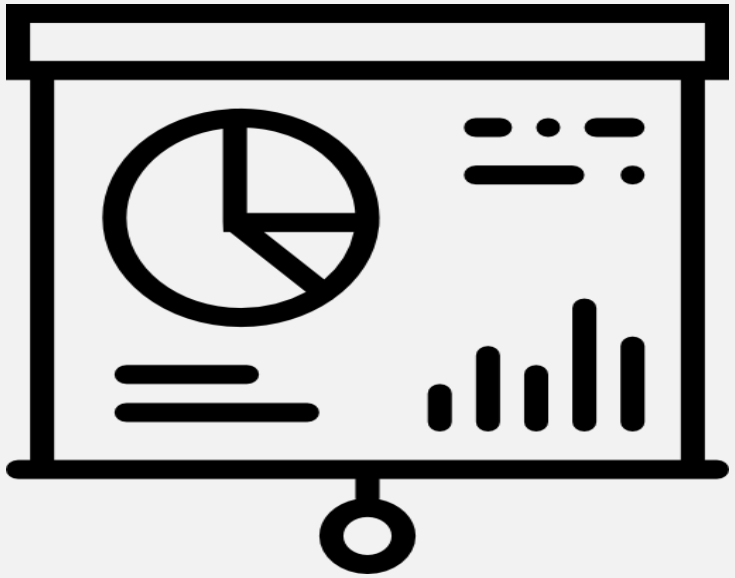
파생변수를 포함시켰을 때의
F-measure가 확연하게 높음을 확인



3

데이터 분석

해석 모형



정교화된 고객 평가

주어진 정보로 연체 위험률 계산

Interpretation

해석모형을 통한 요인분석

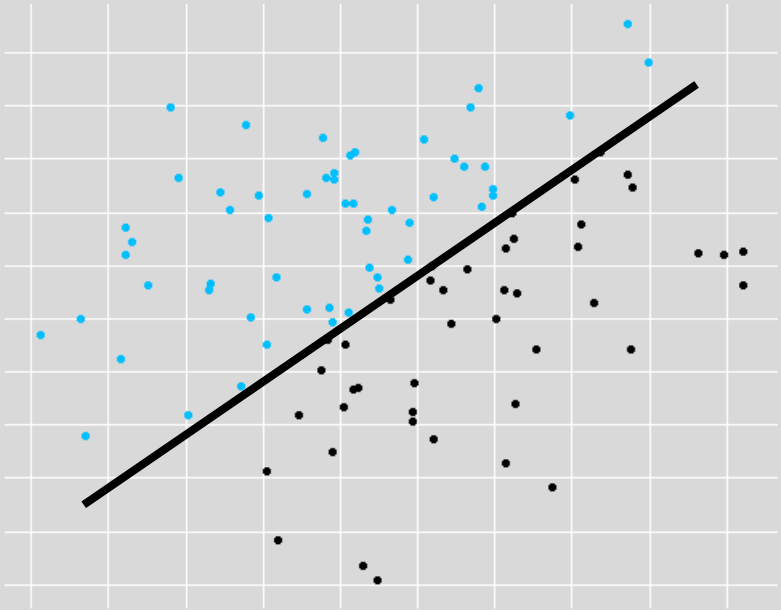
→ 어떤 요인이 유의한가?
→ 요인의 정량적 위험은?

Prediction

예측 모형(머신 러닝)을 통한 예측

→ 모든 변수 사용(유의성이 낮은 요인의 정보량까지 포함)
→ 다양한 변수 조합 및 모형 고려

해석 모형



Generalized Linear Model

관심이 되는 변수의 형태에 크게 구애받지 않고
설명변수와의 선형 관계를 설명할 수 있게 해주는 해석 모형

수준이 0,1 두 개인 변수와 설명변수 사이의 선형관계 규명 가능

일반화 선형모형을 통한
변수 영향력 해석: 변수선택의 필요성

연체위험 = 상환의지 + 상환여력 + 취미 + 근속일수

연체위험 = 상환의지 + 상환여력

영향력이 적은 변수들 존재

→ 영향력이 적은 변수를 제거함으로써 성능 손실을 최소화하는 동시에 해석모델의 효율을 높일 수 있음

일반화 선형모형을 통한 변수 선택 방법1

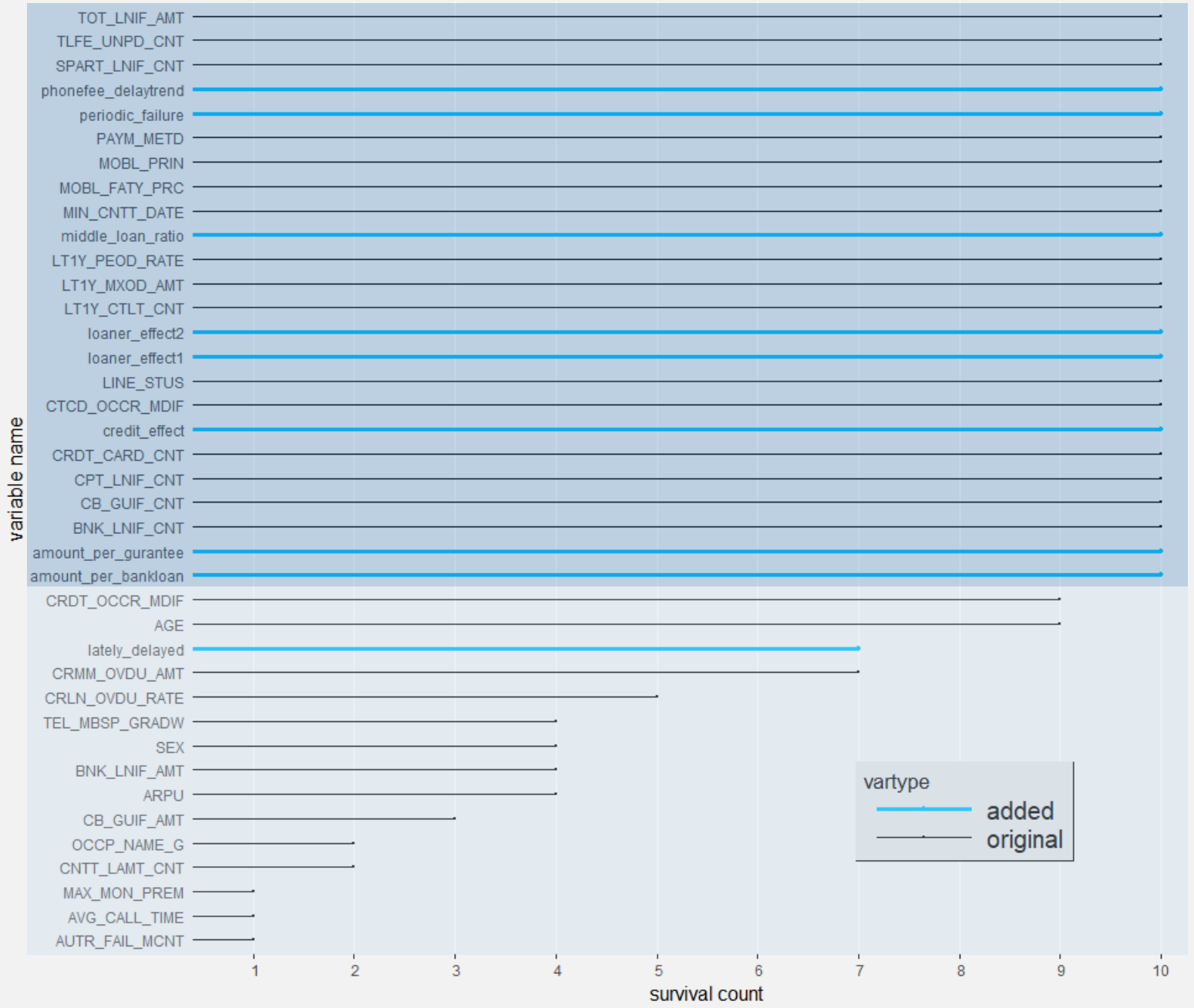
LASSO Regularization

변수를 선택하는 것에 패널티를 줌으로써
적절한 수준에서 변수를 선택할 수 있게 해주는 변수선택법

변수선택에 샘플링에 의한 효과가 미치는 영향 배제 필요
→ 반복 시행을 거쳐 일관되게 선택되는 변수를 최종 선택

일반화 선형모형을 통한 변수 영향력 해석 모형 적합

파생변수를 포함한 데이터에
샘플링을 10번 반복했을 때,
각각의 변수가 LASSO 알고리즘으로
선택된 횟수



일반화 선형모형을 통한 변수 선택 방법2

Boosting

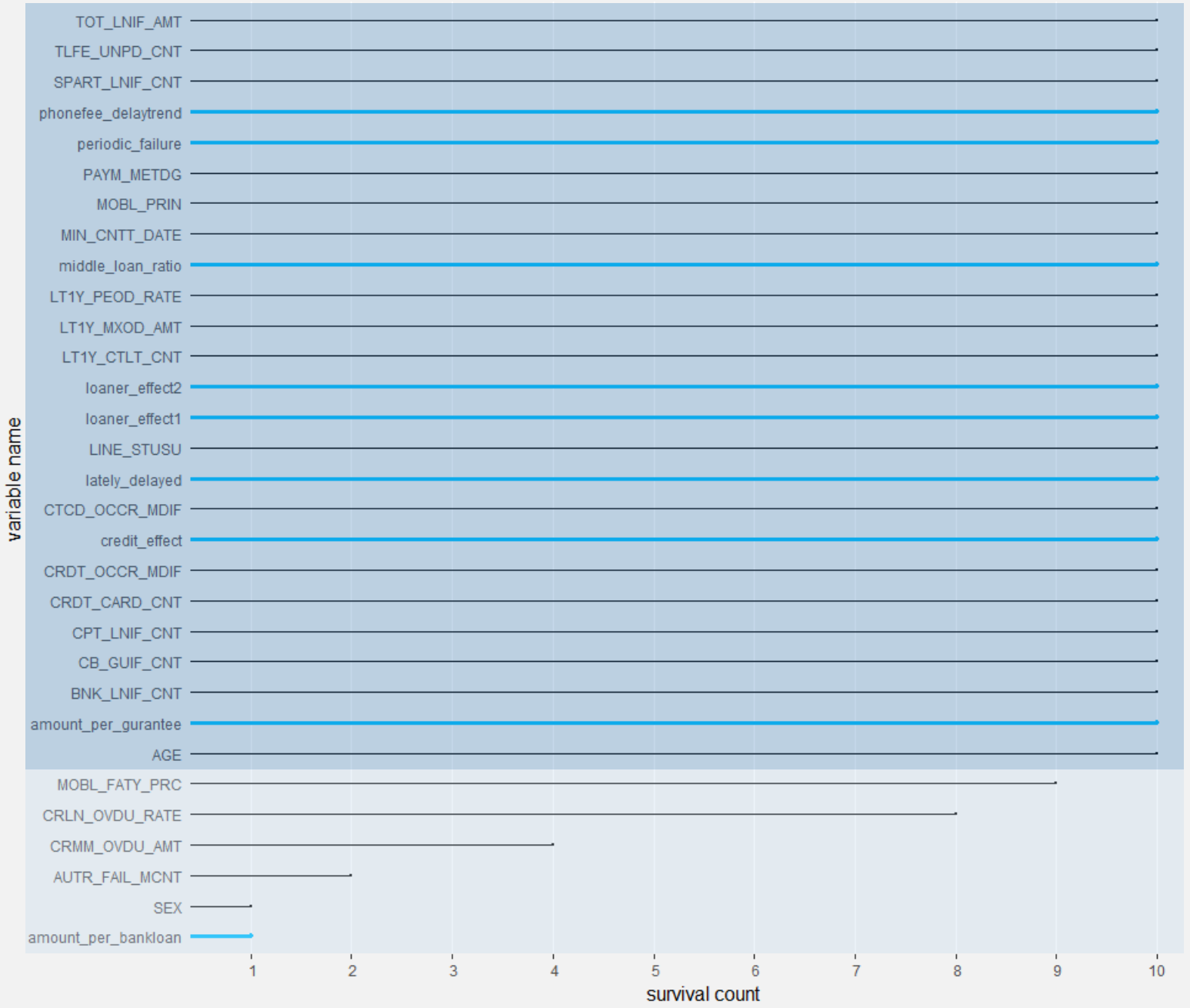
단순하고 약한 모델을 결합해서 보다
정확하고 강력한 모델을 만드는 방식

GLM Boost는 각각의 독립변수 하나의 모델을 단순 모델로 설정
→ 결합을 여러 번 반복하여 최적의 모델로 결합.
→ 최적 모델의 변수를 선택하여 변수선택의 효과 얻음

일반화 선형모형을 통한 변수 영향력 해석 모형 적합

파생변수를 포함한 데이터에
샘플링을 10번 반복했을 때,
각각의 변수가 GLM Boosting에 의해
선택된 횟수

LASSO로 변수선택했을 때에 비해
안정적으로 선택되는 변수의 비중이
커졌음을 알 수 있다.



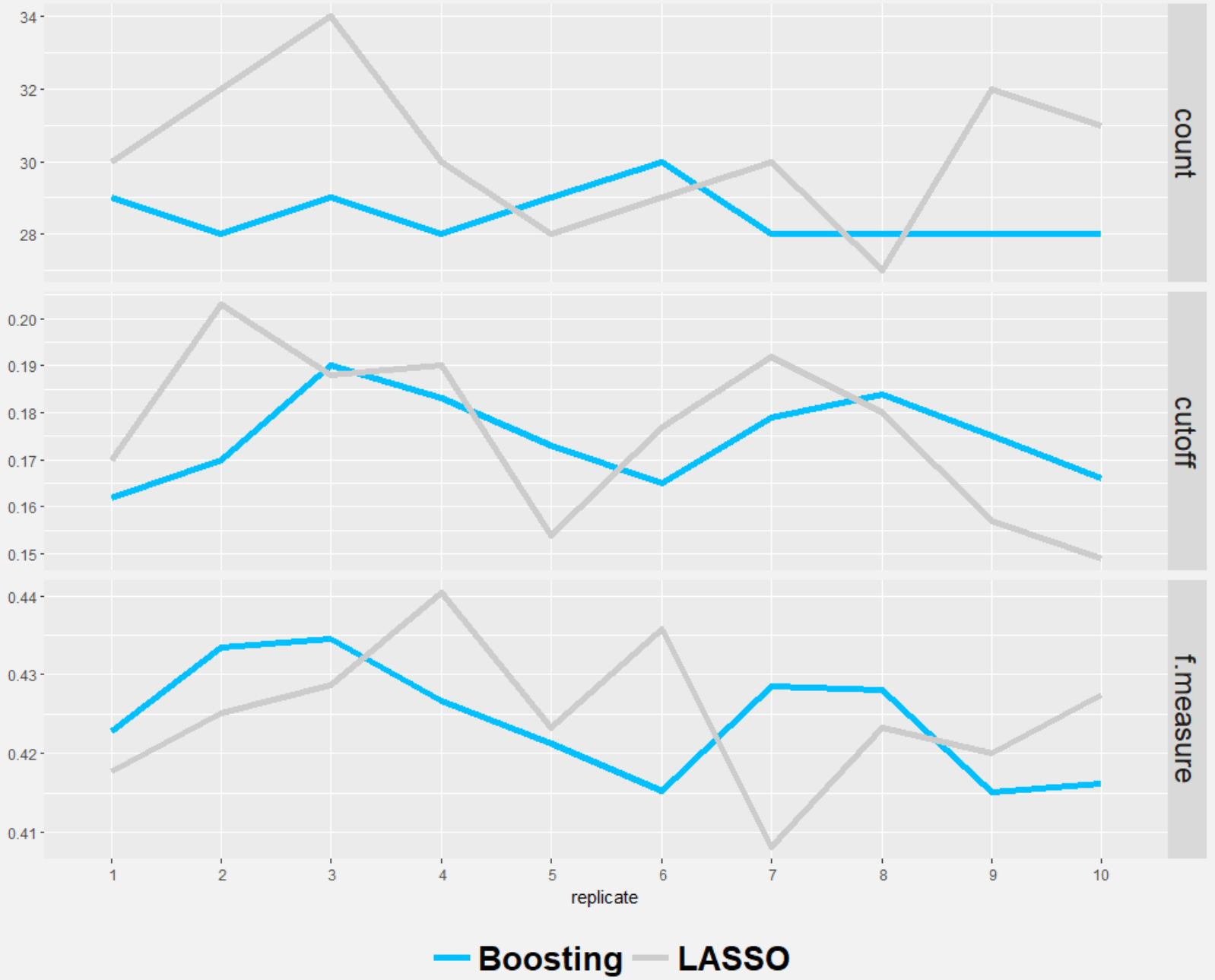
변수선택법 간 비교:
LASSO vs Boosting

파생변수를 추가한 데이터를 가지고
두 가지 변수선택법을 비교

count
선택되는 회차별 변수개수

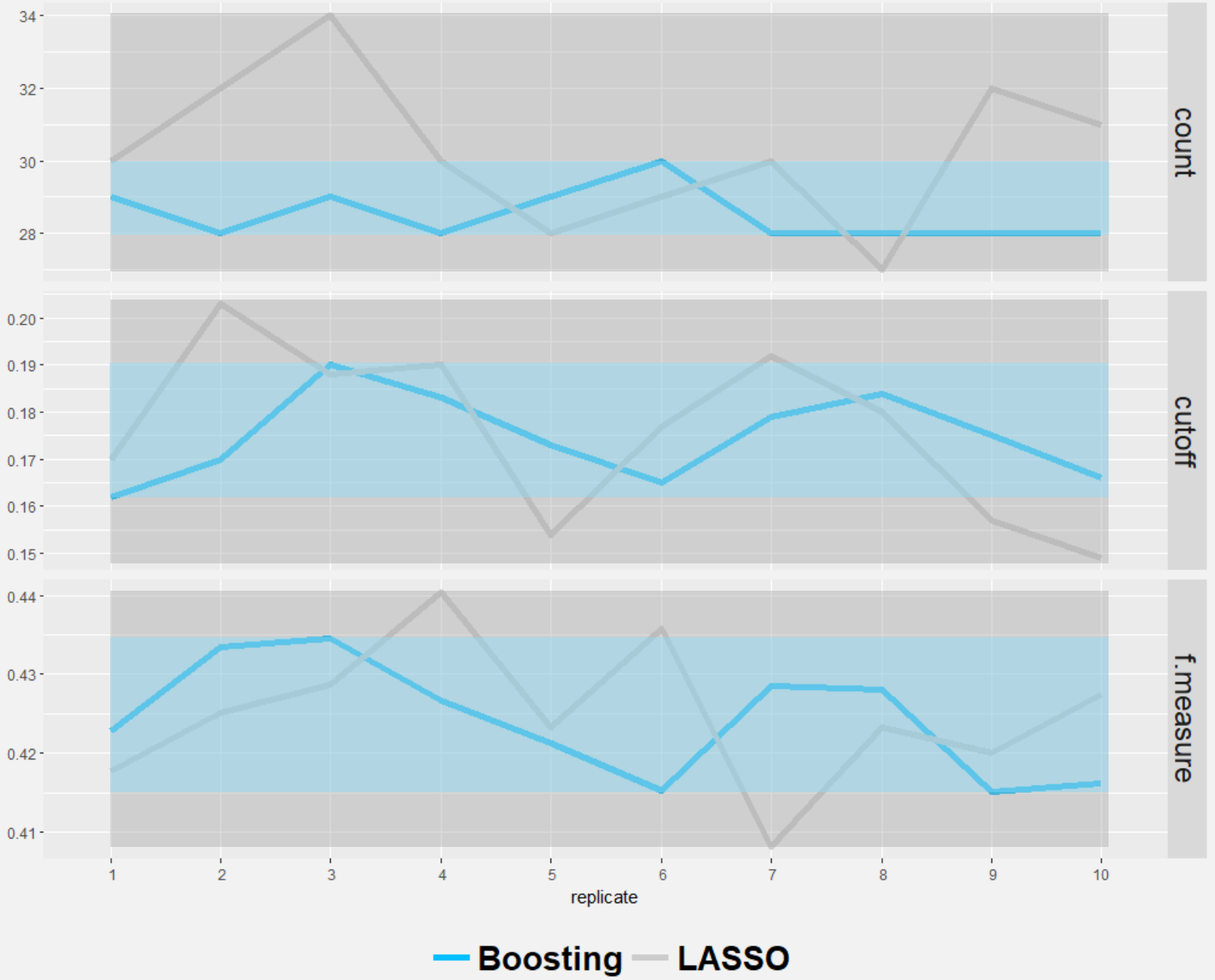
cutoff
F값이 최대화되는 0,1 컷오프

f.measure
회차별 컷오프에서 최대화된 F값

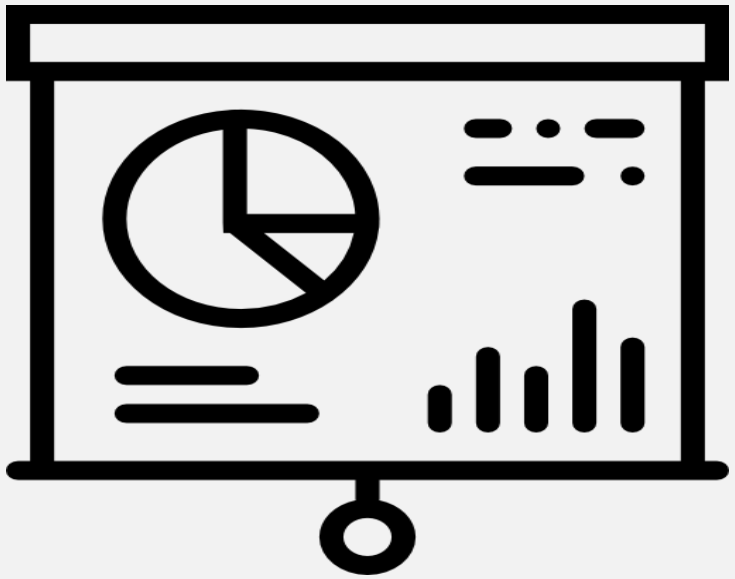


변수선택법 간 비교:
LASSO vs Boosting

세 측정 결과 모두를 통해
Boosting을 통한 변수선택이
샘플링에 의한 변동이 적어서
더 안정적인 변수선택이 가능함을 확인



예측 모형



정교화된 고객 평가

주어진 정보로 연체 위험률 계산

Interpretation

해석모형을 통한 요인분석

→ 어떤 요인이 유의한가?
→ 요인의 정량적 위험은?

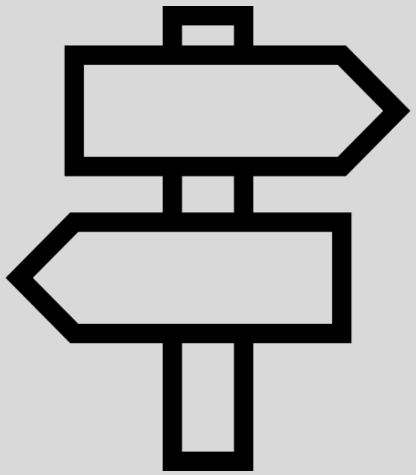
Prediction

예측 모형(머신 러닝)을 통한 예측

→ 모든 변수 사용(유의성이 낮은 요인의 정보량까지 포함)
→ 다양한 변수 조합 및 모델 고려

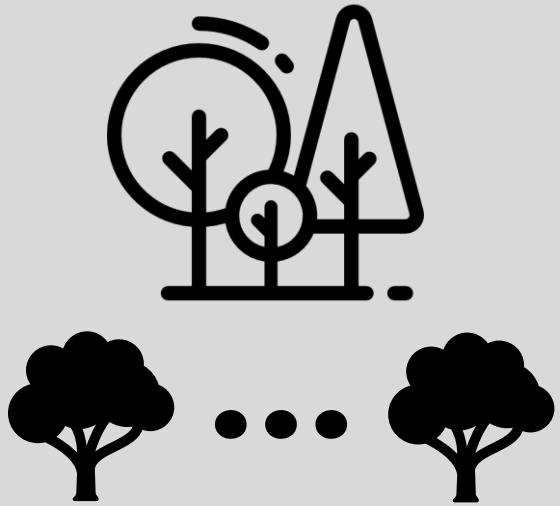
예측 모형

SVM



데이터에 적합한 분류기준선으로
반응을 이진분류하는 분류기

Random Forest



여러 개의 의사결정나무의
예측결과를 종합하는 분류기

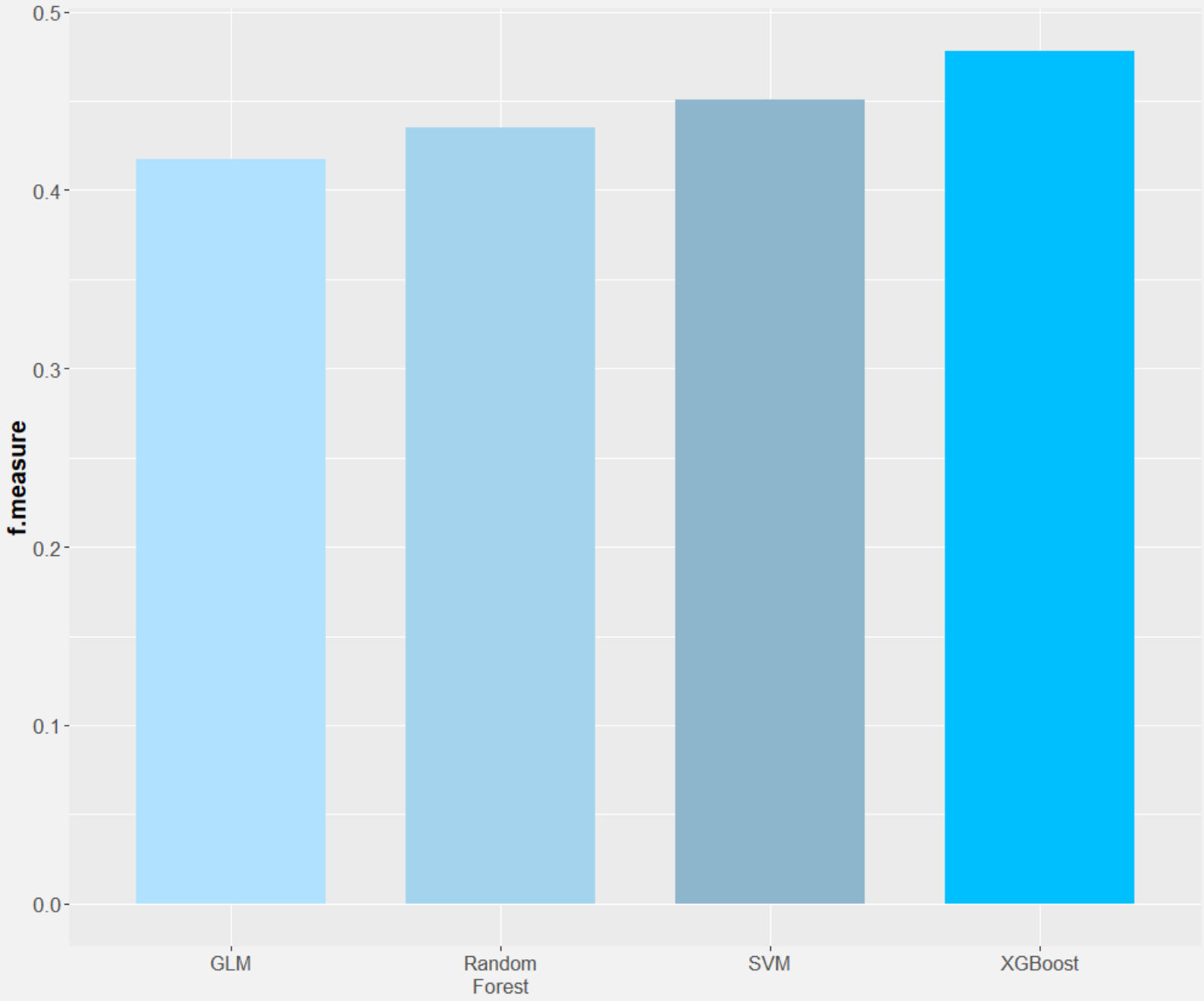
XGBoost

에러가 감소하도록 모델링을 반복해
얻은 최적의 모델을 사용하는 분류기

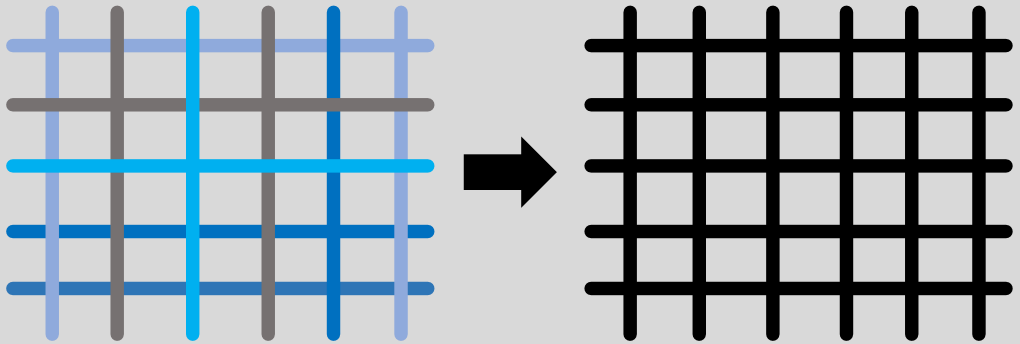
예측 모델간 단일 성능 비교

각 예측 모델들을 반복 적합하면서
모형별로 옵션을 계속 조정한 결과:

평균적으로 XGBoost가 최우수 분류기



앙상블 기법



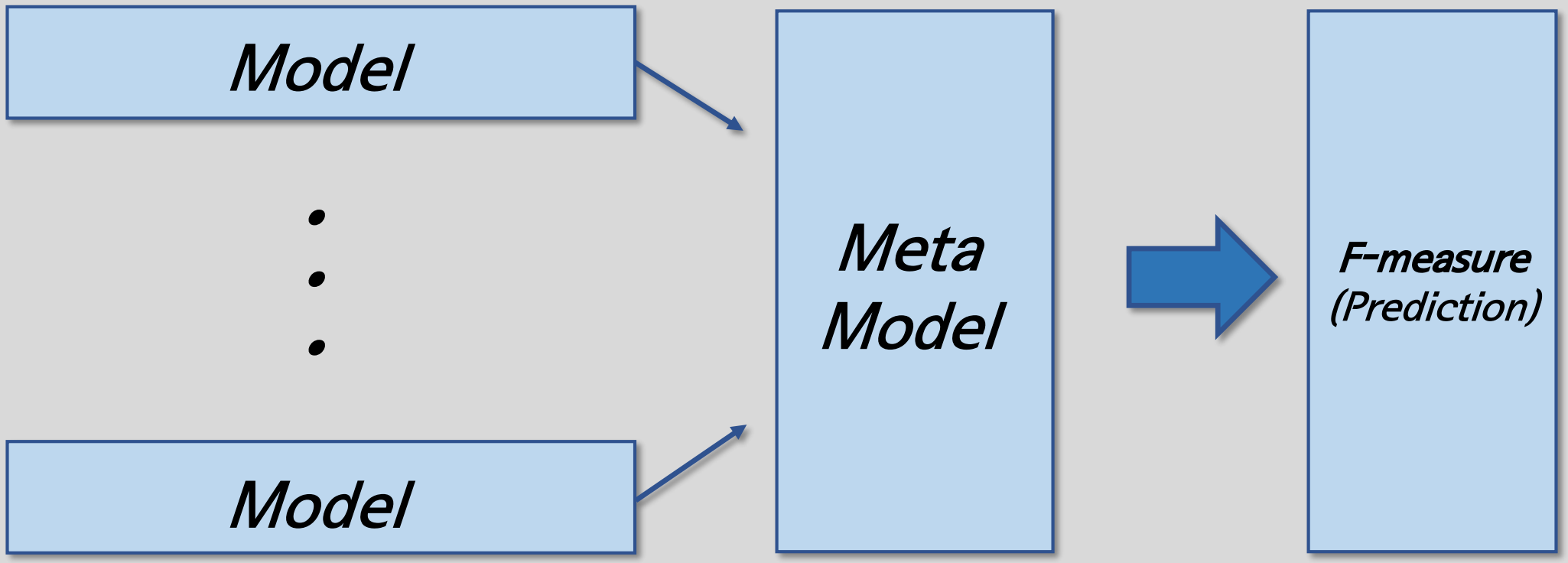
Ensemble method

여러가지 모델의 결과값들에 가중치를 부여하여
이를 종합해 예측하는 방법

→ 샘플링에 의한 단일 모델의 변동을
다른 모델이 완충하는 효과

→ 모델 간 상호보완을 통해 최종적으로 예측력 상승

앙상블 기법: 예측과정

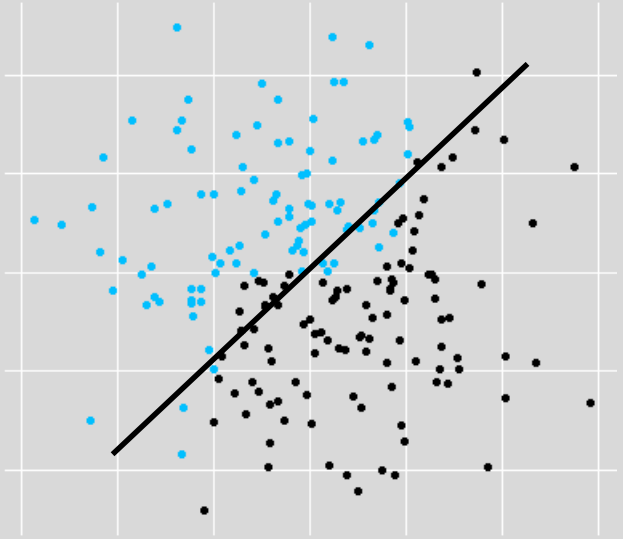


앙상블 기법: 종합하는 방법

Voting

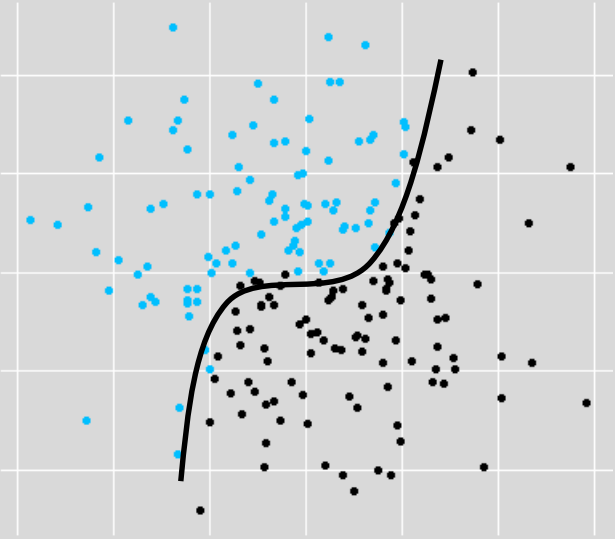
- ✓ 1차 모델 예측결과 단순 취합
- ✓ 성능 좋은 모델 과소 평가 우려

GLM



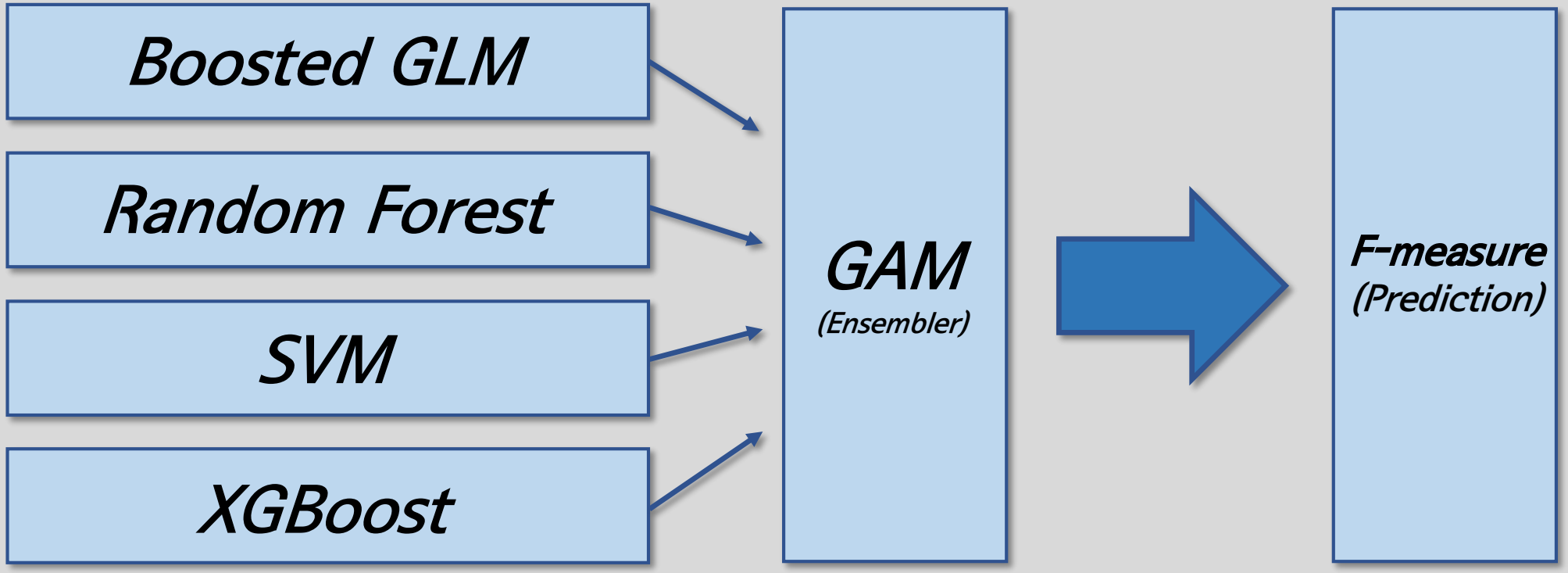
- ✓ 가중치 부여해 모델 별 특성 반영
- ✓ 선형적인 경우에만 효과적

GAM

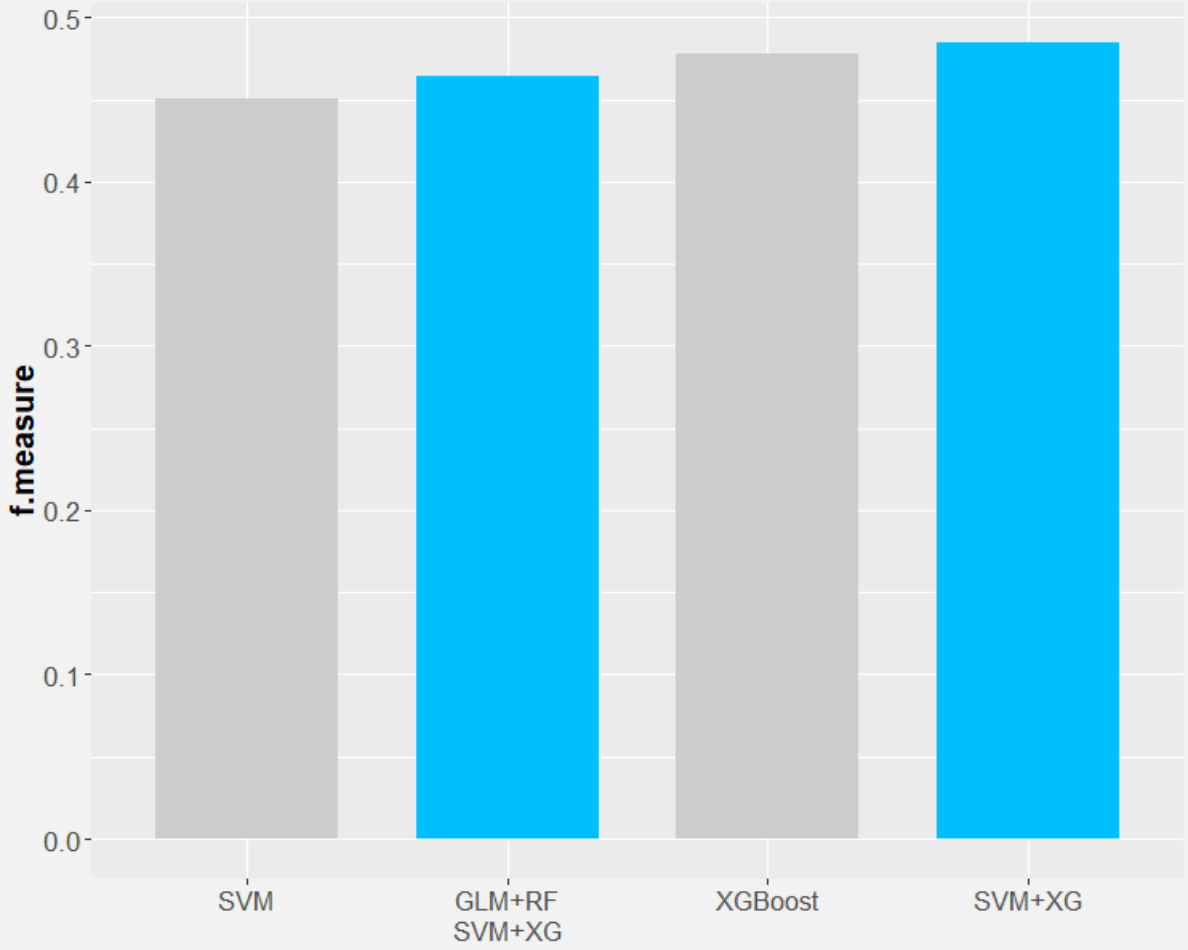


- ✓ 가중치 부여해 모델 별 특성 반영
- ✓ 비선형적 경향성 포착가능

앙상블 기법: 1차 모델 선정



앙상블 기법: 1차 모델 선정

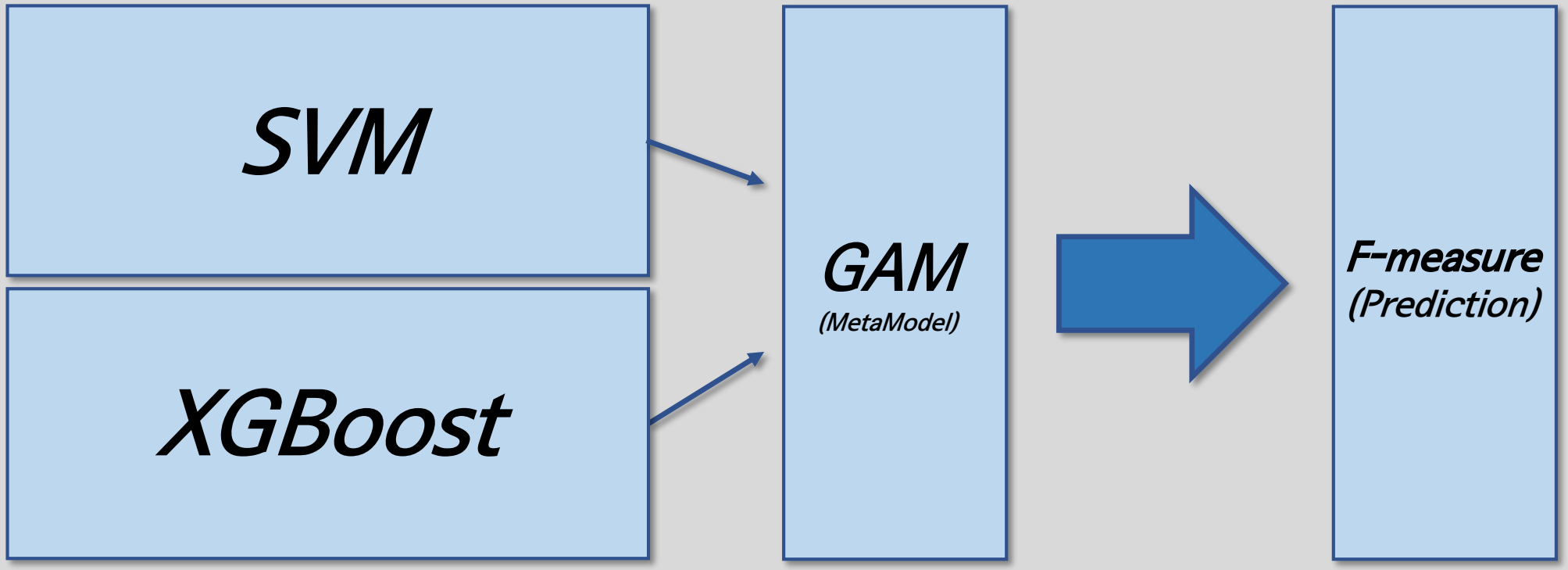


성능 낮은 모델의 보정효과 미미

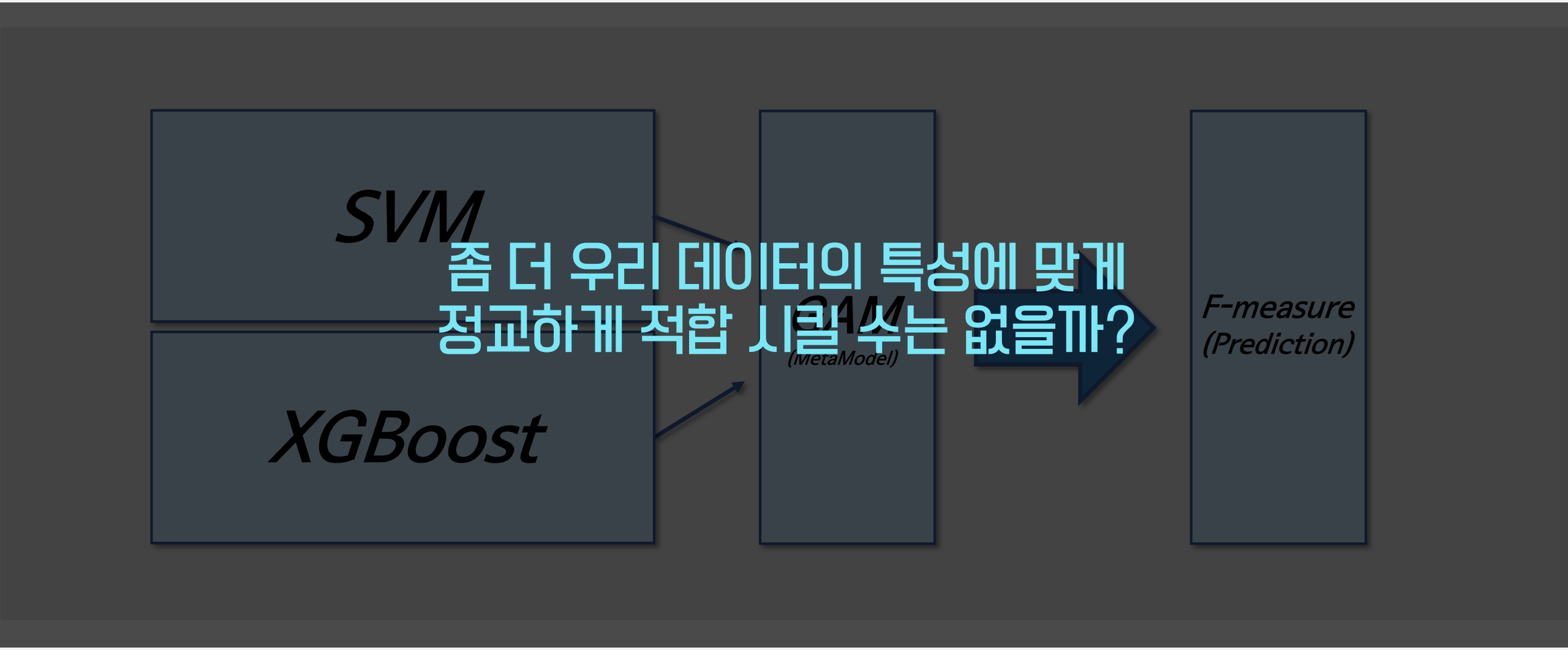
성능이 낮은 쪽으로 하향 평준화

성능 낮은 GLM, RandomForest 제외

1차모델 선정결과

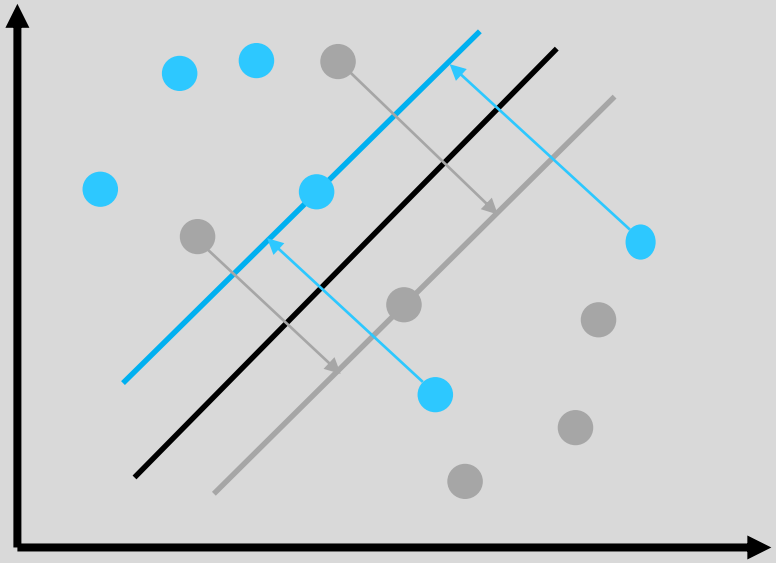


1차모델 선정결과

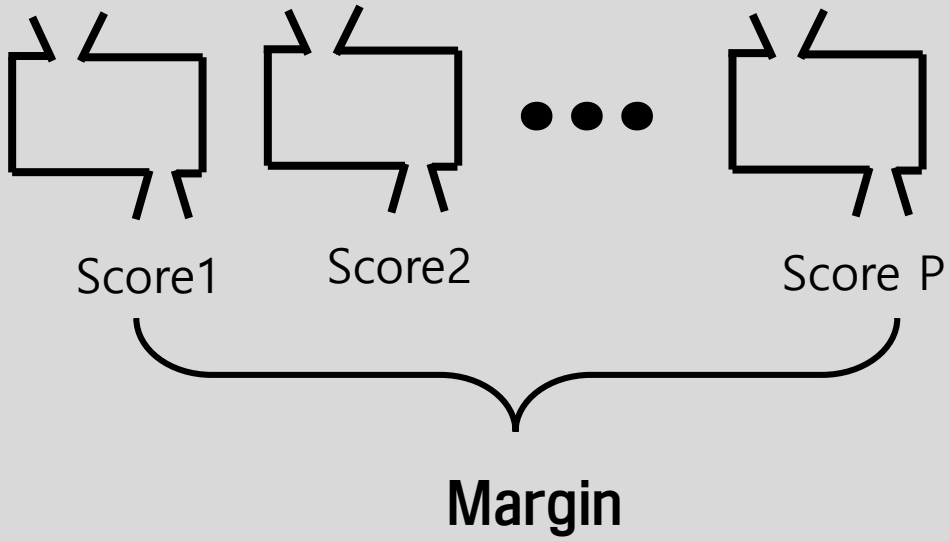

SVM
XGBoost
GAM
(MetaModel)
좀 더 우리 데이터의 특성에 맞게
정교하게 적합 시킬 수는 없을까?
F-measure
(Prediction)

앙상블 기법: 2차 모델 선정

SVM Distance

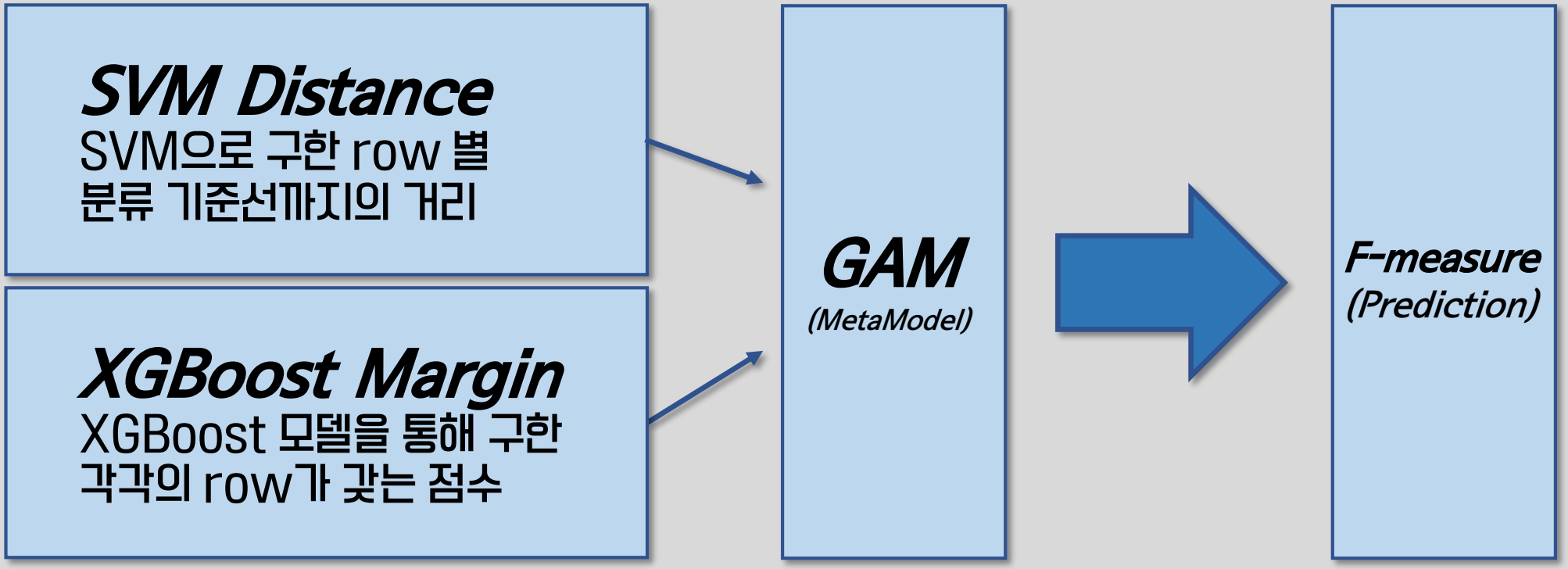


XGBoost Margin



모델 자체의 특성을 반영하기 위해
확률 값 대신, 모델 고유의 값 자체를 사용

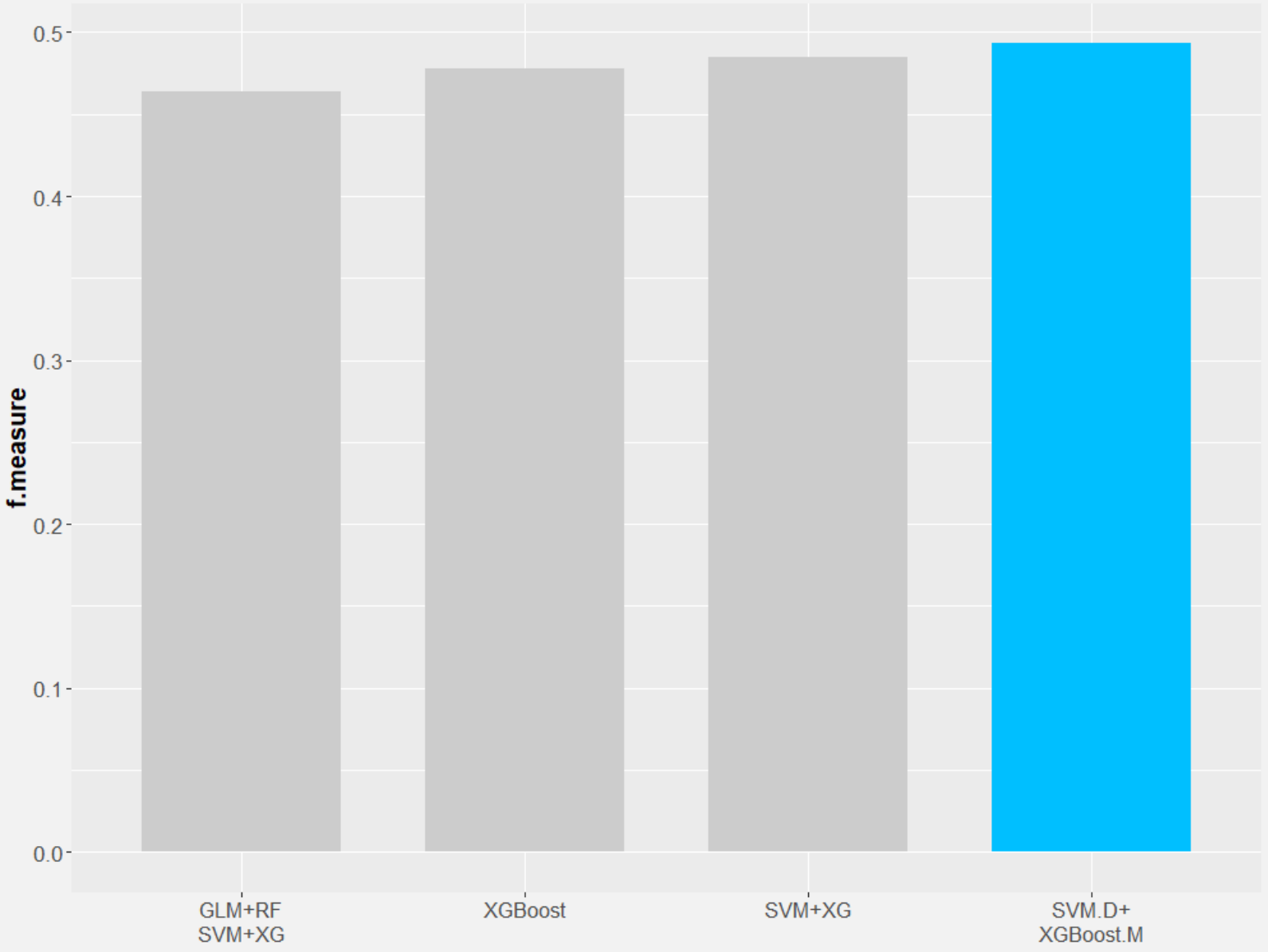
예측모델 선정결과


SVM Distance
SVM으로 구한 row 별
분류 기준선까지의 거리
XGBoost Margin
XGBoost 모델을 통해 구한
각각의 row가 갖는 점수
GAM
(MetaModel)
F-measure
(Prediction)

예측 모델 별 결과 비교

단일 모델 중 최고였던 XGBoost와
3가지 경우의 앙상블 모형의
F-measure 비교

F-measure가 가장 우수한
SVM.D + XGBoost.M 앙상블 선택

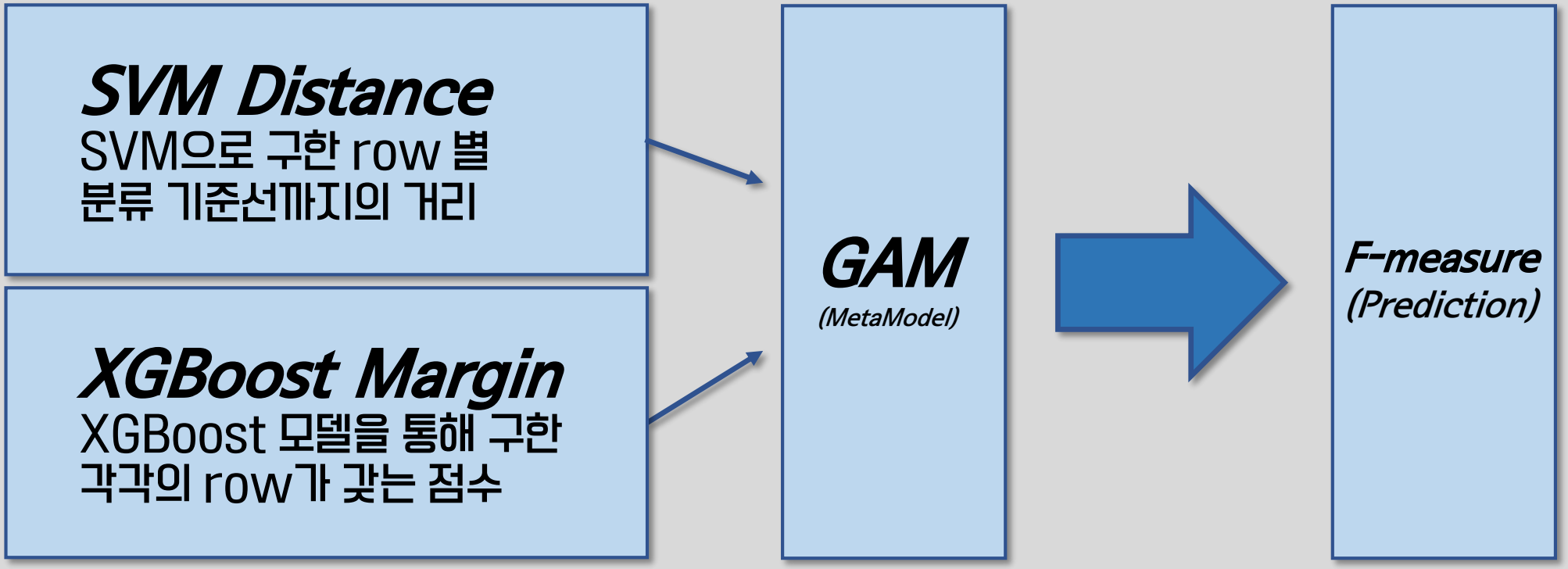


모두 앙상블 < XGBoost 단일 < SVM, XG 앙상블 < SVM.D, XG.M 앙상블

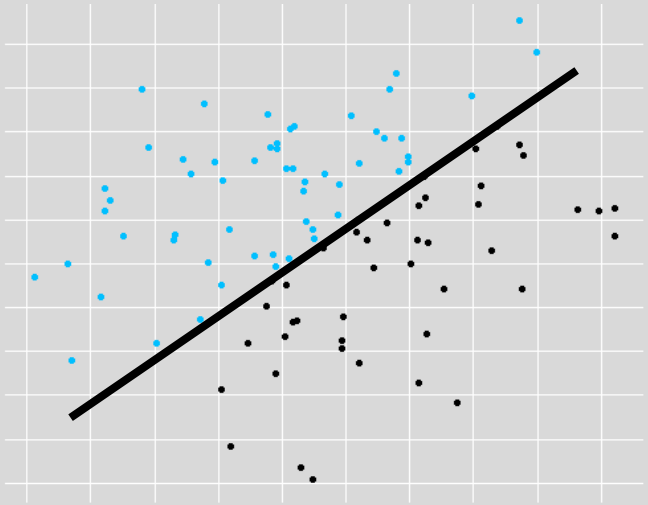
4

결론 및 의의

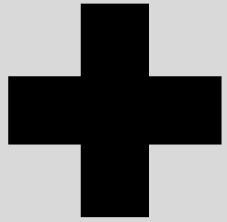
예측모델 선정결과

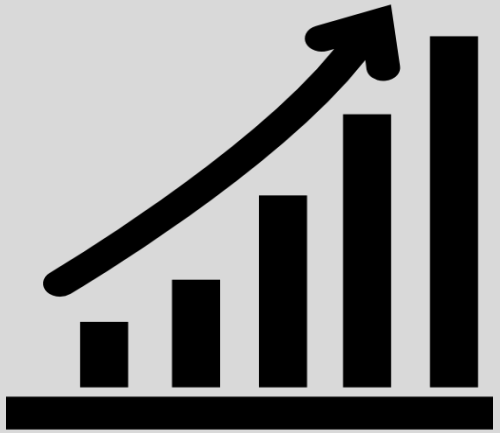

SVM Distance
SVM으로 구한 row 별
분류 기준선까지의 거리
XGBoost Margin
XGBoost 모델을 통해 구한
각각의 row가 갖는 점수
GAM
(MetaModel)
F-measure
(Prediction)

해석모델 선정결과



GLM



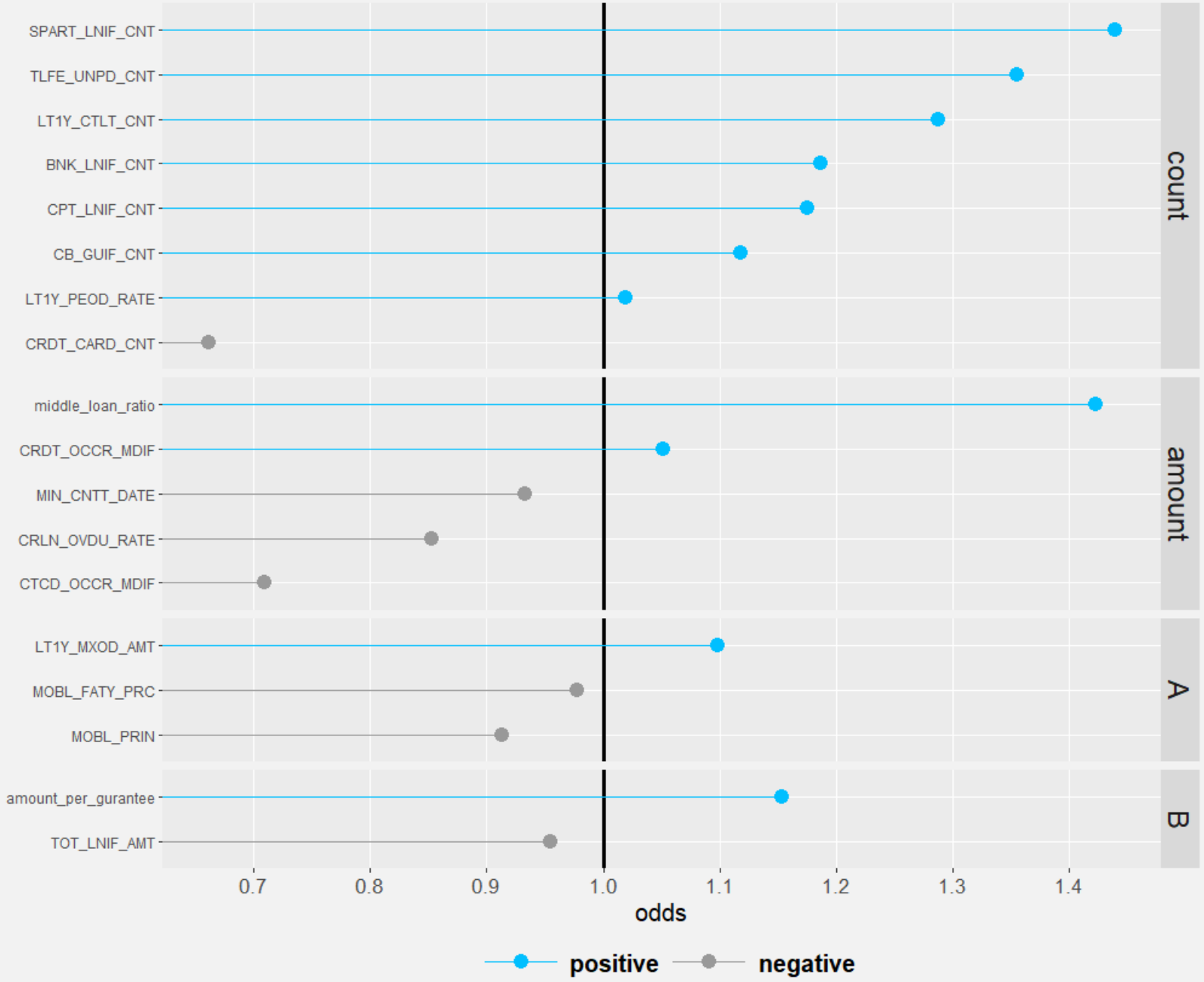


Boosting

Boosting을 사용한 GLM 해석모델 구축

변수선택 및 계수추정 결과 : 숫자형 변수를 중심으로

count의 단위가 1,
amount의 단위가 100,
A, B의 단위가 각각 10만원, 1억일 때,
각 변수가 1단위 증가할 때의 연체위험 변화



변수선택 및 계수추정 결과 : 파생변수 해석을 중심으로

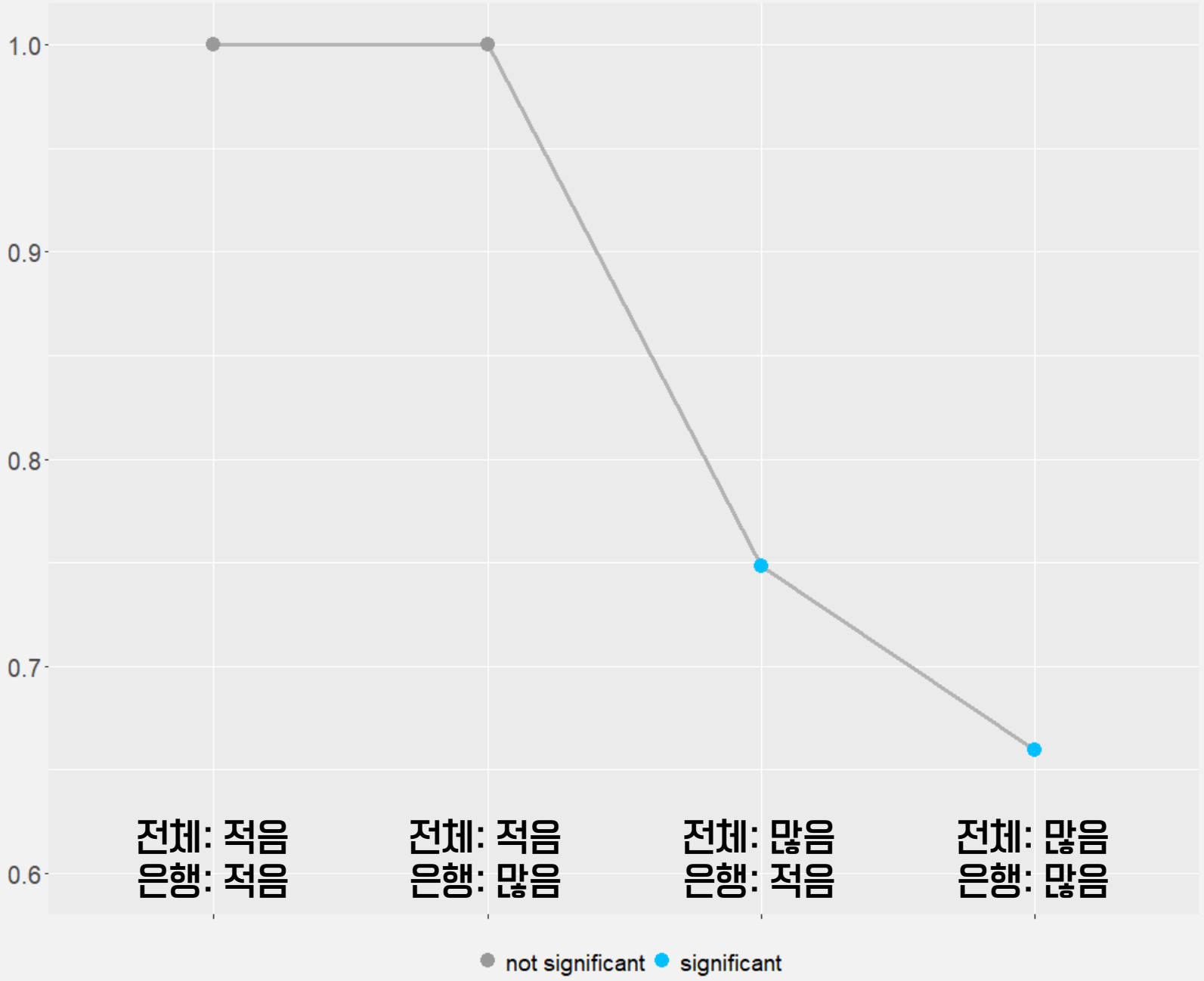
대출금액 비교

대출 총액이 낮은 집단이 대출 총액이 높은 집단보다 연체 위험 평균 0.66배

대출금액(특히 은행) 많으면 연체 위험 감소

관련 변수

- 대출정보 현재 총 금액
- 대출정보 현재 총 금액(은행)



변수선택 및 계수추정 결과 : 파생변수 해석을 중심으로

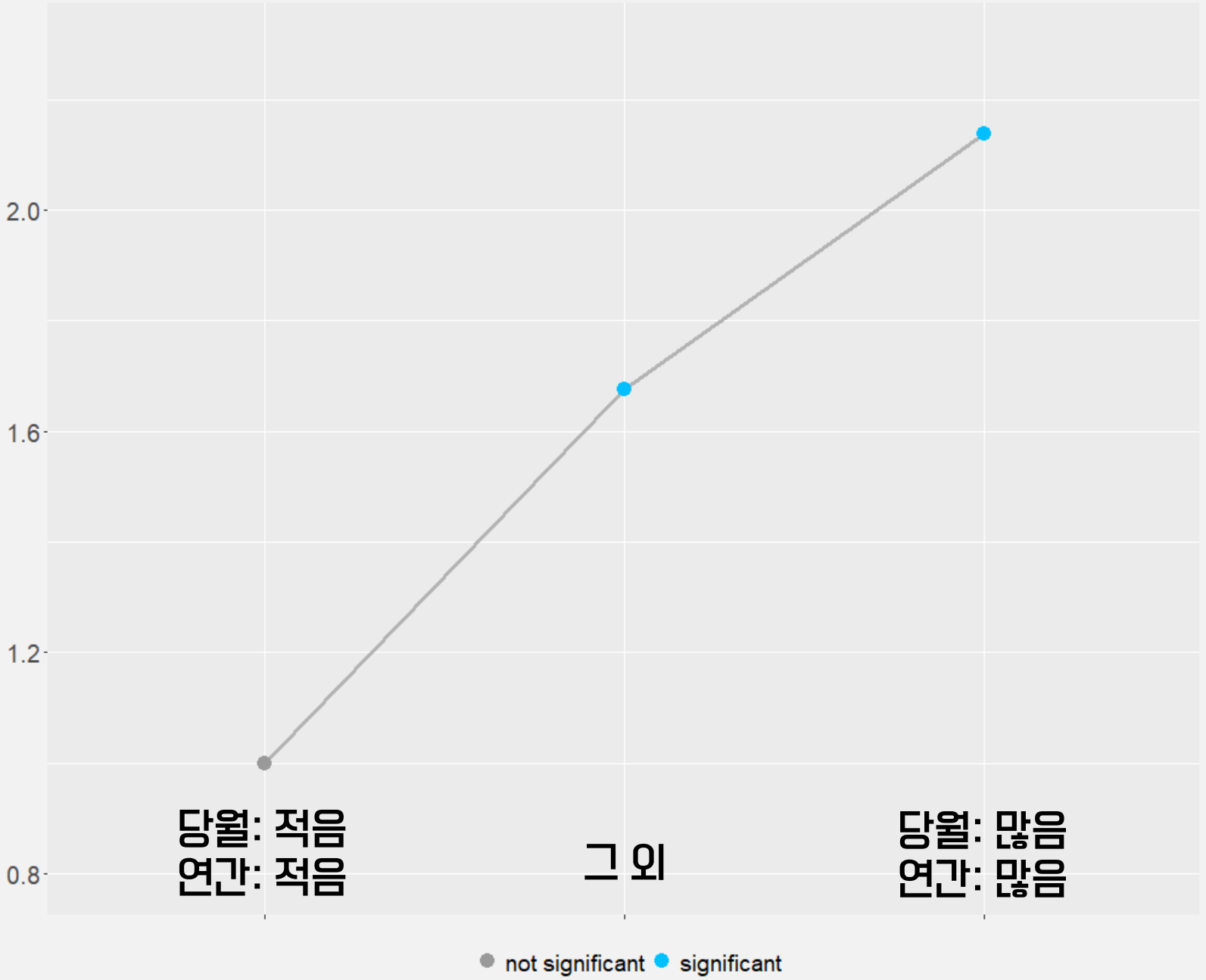
통신비 연체경향

통신비 납부가 불성실한 사람이 성실한 사람보다 연체 위험 평균 2.13배

통신비 납부 태도가 불성실할수록 연체 위험 증가

관련 변수

- 통신비 연간 최대 연체 금액
- 통신비 당월 연체 금액



변수선택 및 계수추정 결과 : 파생변수 해석을 중심으로

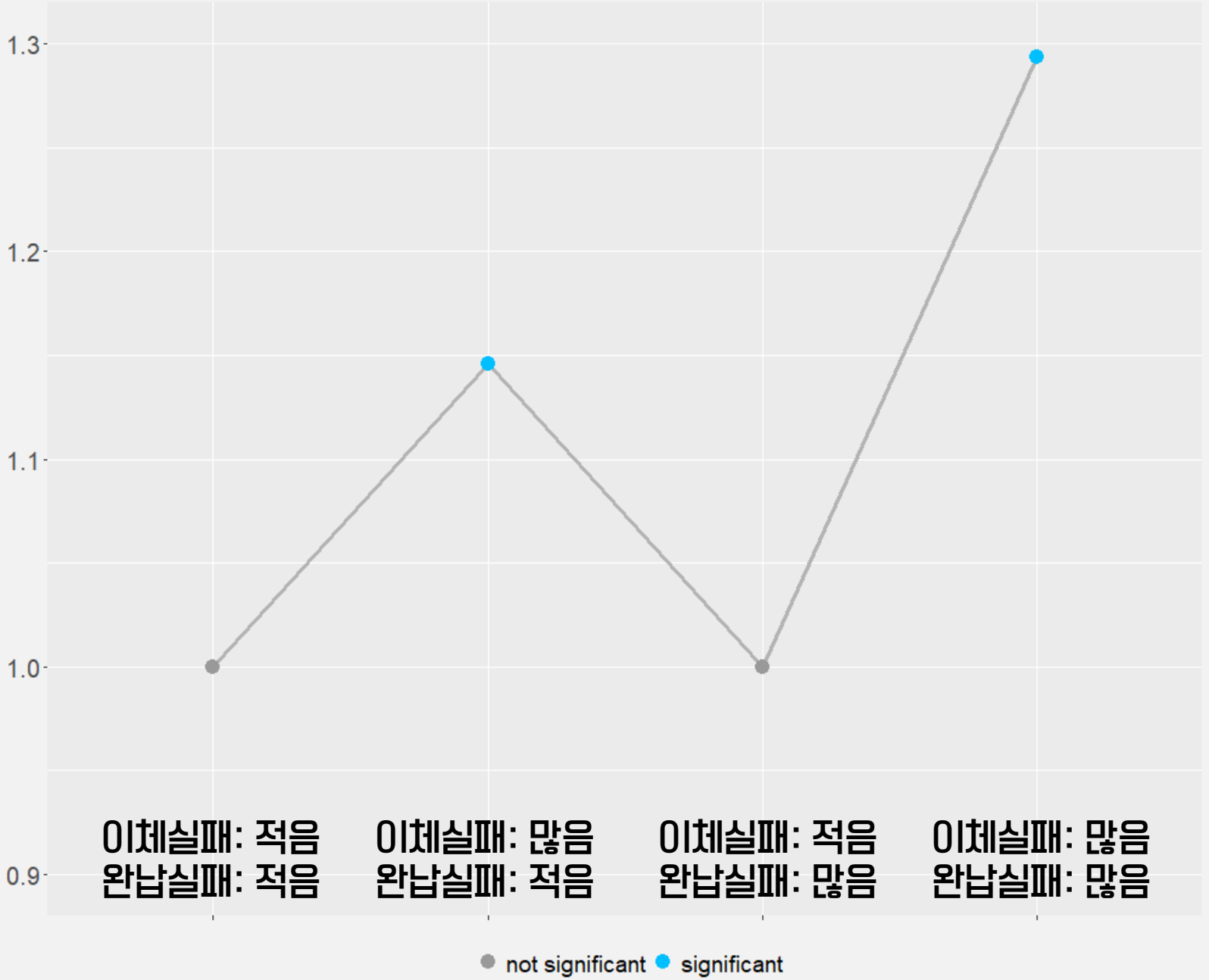
보험료 정기납부 연체효과

납부 성실 그룹에서 납부 불성실 그룹으로 변할 때 연체 위험 평균 1.29배

보험료 납부 태도가 불성실할수록 연체 위험 증가

관련 변수

- 보험료 자동이체 실패 월 수
- 보험료 완납경험 횟수



변수선택 및 계수추정 결과 : 파생변수 해석을 중심으로

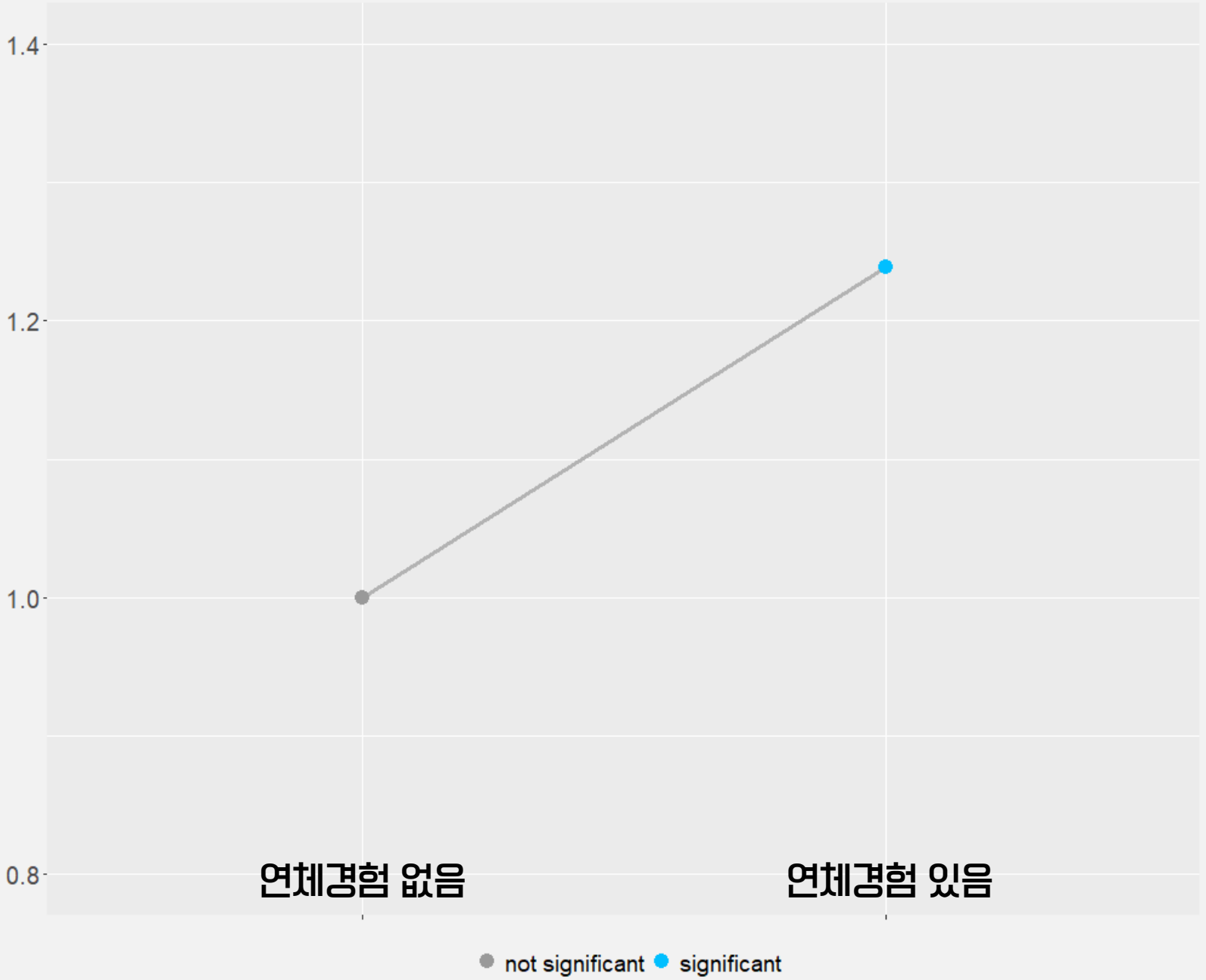
한달 내 연체경험유무(보험&통신)

한달 내 연체경험이 있을 때, 그렇지 않을 때보다 연체 위험 평균 1.24배

최근(한달)에 연체 경험하면 연체 위험 증가

관련 변수

- 통신비 당월 연체 금액
- 근 30일 기준 한화생명 대출 연체율



분석 의의

1. 3사 결합데이터를 활용, 대출정보가 부족한 고객에 대한 연체 위험률 계산

2. 비식별데이터 분석 방향 제시

3. 대출 연체 요인의 영향을 정량적으로 수치화

conclusion