개인신용평가 모델링 과제

2019.9.26

KCB 인턴 4조

목 차

Ⅰ. 변수 소개 – 데이터 구조

차주데이터		계좌데이터
대출이력 있음 (310,434)	+	대출이력 있음
대출이력 없음 (533,180)		

차주데이터	기준월	시점	
Train + Validation	201609 201612 201703	과거시점	
	201706	기준시점	기불량자(해당시점 대출, 카드 연체자) 제외 (5,854)
Test	201709 201712 201803	과거시점	
	201806	기준시점	기불량자(해당시점 대출, 카드 연체자) 제외 (6,174)

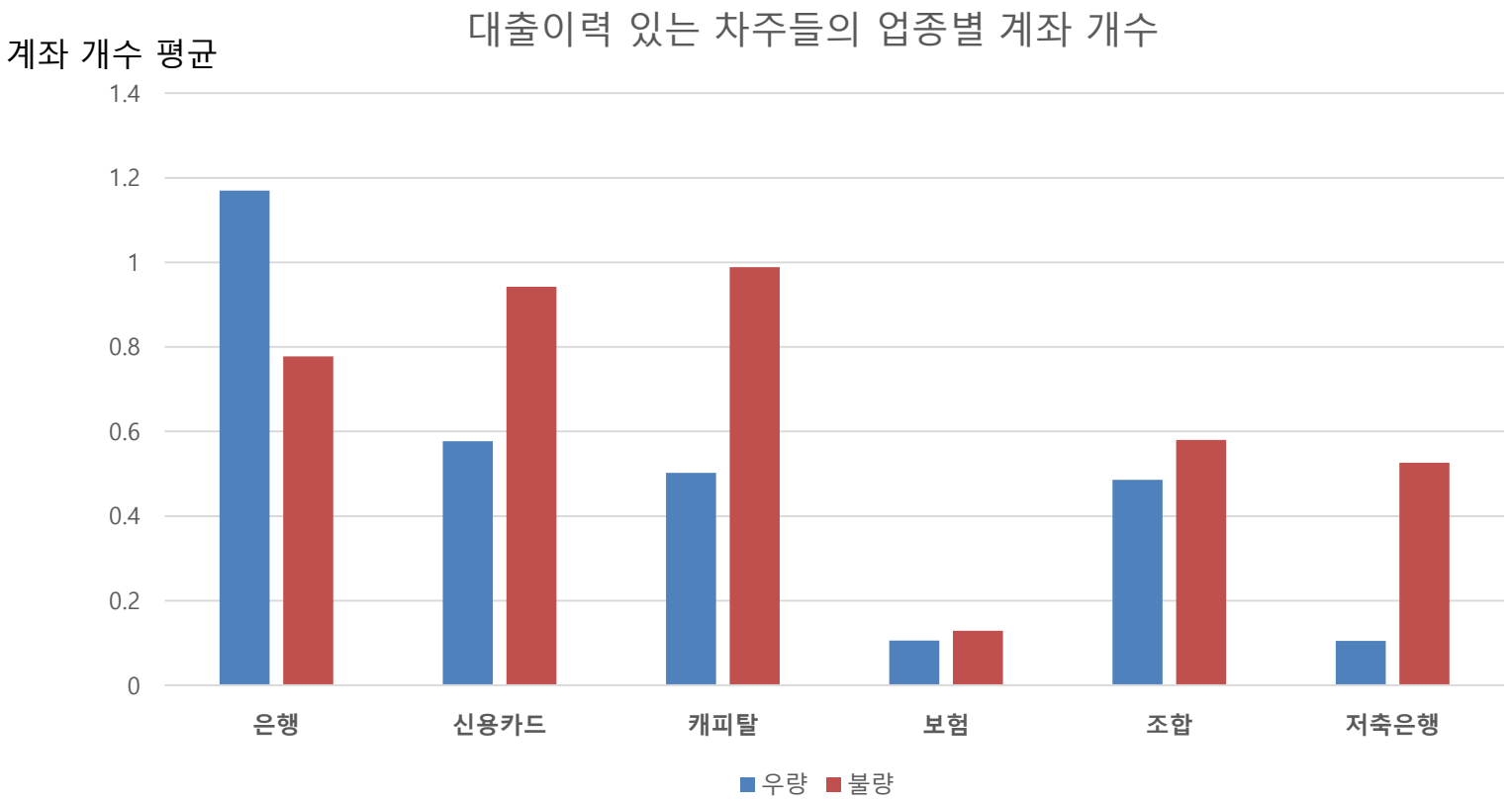
Ⅰ. 변수 소개 – 차주단위 데이터

변수명	변수설명	변수유형
AGE	연령대(20대, 30대,..., 70대 이상)	범주형
INCOME	연 소득	수치형
DTI	총부채상환비율	수치형
JOB	직업(급여소득자, 자영업자, 무직, 정보 없음)	범주형
HOME	자택광역시(강원,...,충북)	범주형
SEX_CD	성별(남자, 여자)	범주형
LN_FLAG	대출보유여부(1:보유, 0:미보유)	범주형
CARD_FLAG	카드보유여부(1:보유, 0:미보유)	범주형
CNT_CARD_USE	이용 신용카드 기관수	수치형
TOT_USE_AMT	신용카드 총 이용금액	수치형
DLQ_N1YE_FLAG	우·불량여부(1:불량(21,389), 0:우량(822,225)), TARGET	범주형

Ⅰ. 변수 소개 – 파생변수

변수명	변수설명	변수유형
JBLN0000	최근 1년 간 비 은행 비 주택담보대출 이용 대출 계좌 수	수치형
JBLN0002	최근 1년 간 비 은행 주택담보대출 이용 대출 계좌 수	수치형
JBLN0100	최근 1년 간 은행 비 주택담보대출 이용 대출 계좌 수	수치형
JBLN0102	최근 1년 간 은행 주택담보대출 이용 대출 계좌 수	수치형
MRTY_LEFT	기준시점 유효한 계좌들의 만기까지 남은 기간	수치형
INCOME_SD	최근 1년 간 연 소득 표준편차	수치형
CARD_MEAN_CNT	최근 1년 간 신용카드 평균 이용기관 수(0, 1이상으로 구분)	범주형
INCOME_CARD	최근 1년 간 월 소득 대비 신용카드 평균 이용금액	수치형
INCOME_DELAY_AMOUNT	최근 1년 간 연소득대비 연체잔액	수치형
CARD_DELAY_SCORE	카드연체점수. 연체시기가 기준시점에 가까울수록 높은 가중치 부여	수치형
ACCT_DELAY_SCORE	대출연체점수. 연체시기가 기준시점에 가까울수록 높은 가중치 부여	수치형

1. 최근 1년간 업종별 대출 계좌 개수 (integer)

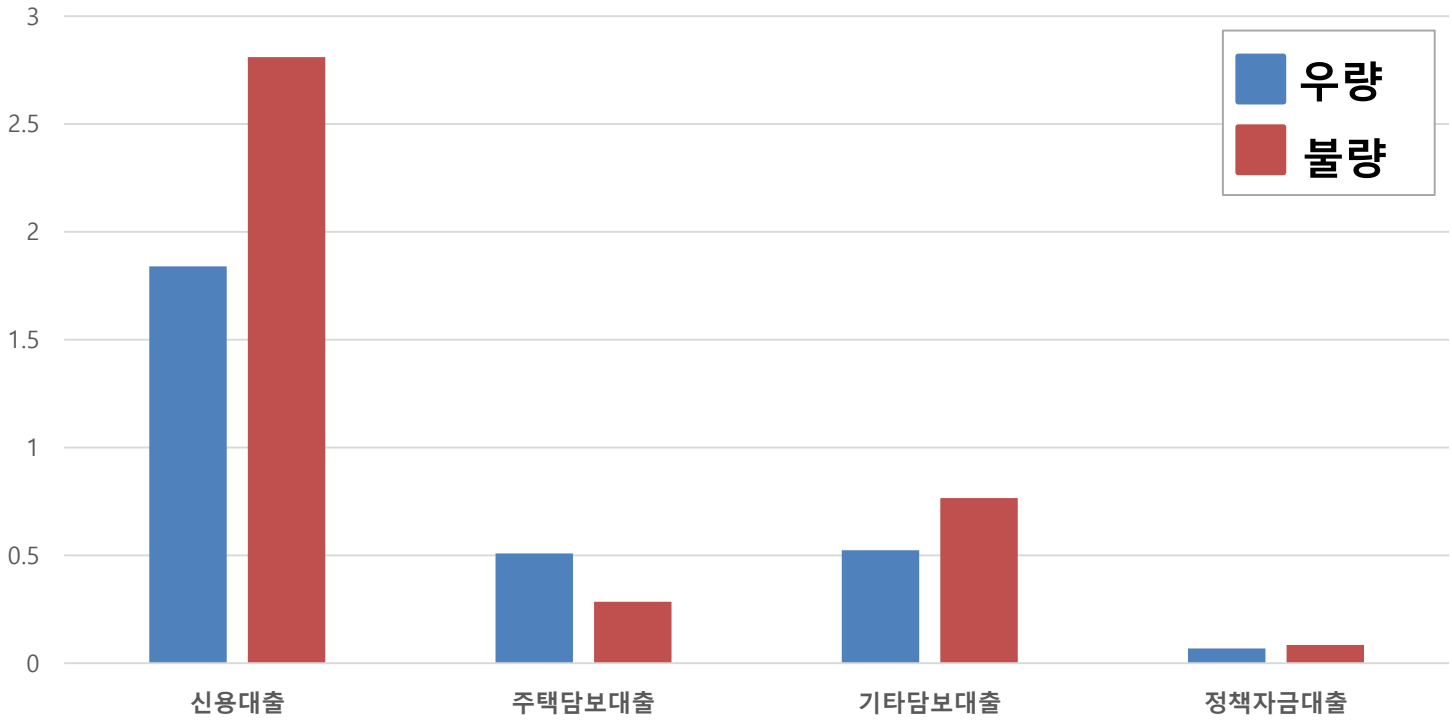


- 은행의 경우에서만 우량 차주 계좌 수 > 불량 차주 계좌 수
- 업종을 은행과 비 은행으로 통합

2. 최근 1년간 상품별 대출 계좌 개수 (integer)

계좌 개수 평균

대출이력 있는 차주들의 상품별 계좌 개수



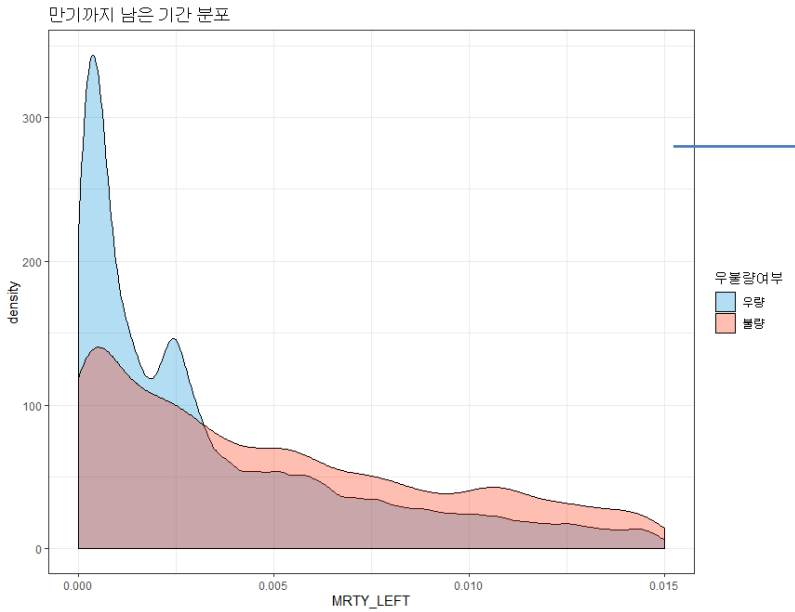
- 주택담보대출의 경우에서만 우량 차주 계좌 수 > 불량 차주 계좌 수
- 주택담보대출/비주택담보대출로 변수 통합

3. 만기까지 남은 기간 (numeric)

$$\text{MRTY\_LEFT}_i := \sum(\text{업종별 만기까지 남은 기간} \times \text{업종별 불량률})$$

각 계좌의 만기까지 남은 기간(분기) 계산

- 상품별로 남은 기간 표준화: 신용대출에 비해 주택담보대출이 대출기간이 매우 긴 편
- 불량률이 높은 업종의 대출 계좌 만기가 많이 남아있을수록 MRTY_LEFT가 큼



- **불량 차주일수록 MRTY_LEFT가 전반적으로 높음**

4. 최근 1년 간 연 소득 표준편차

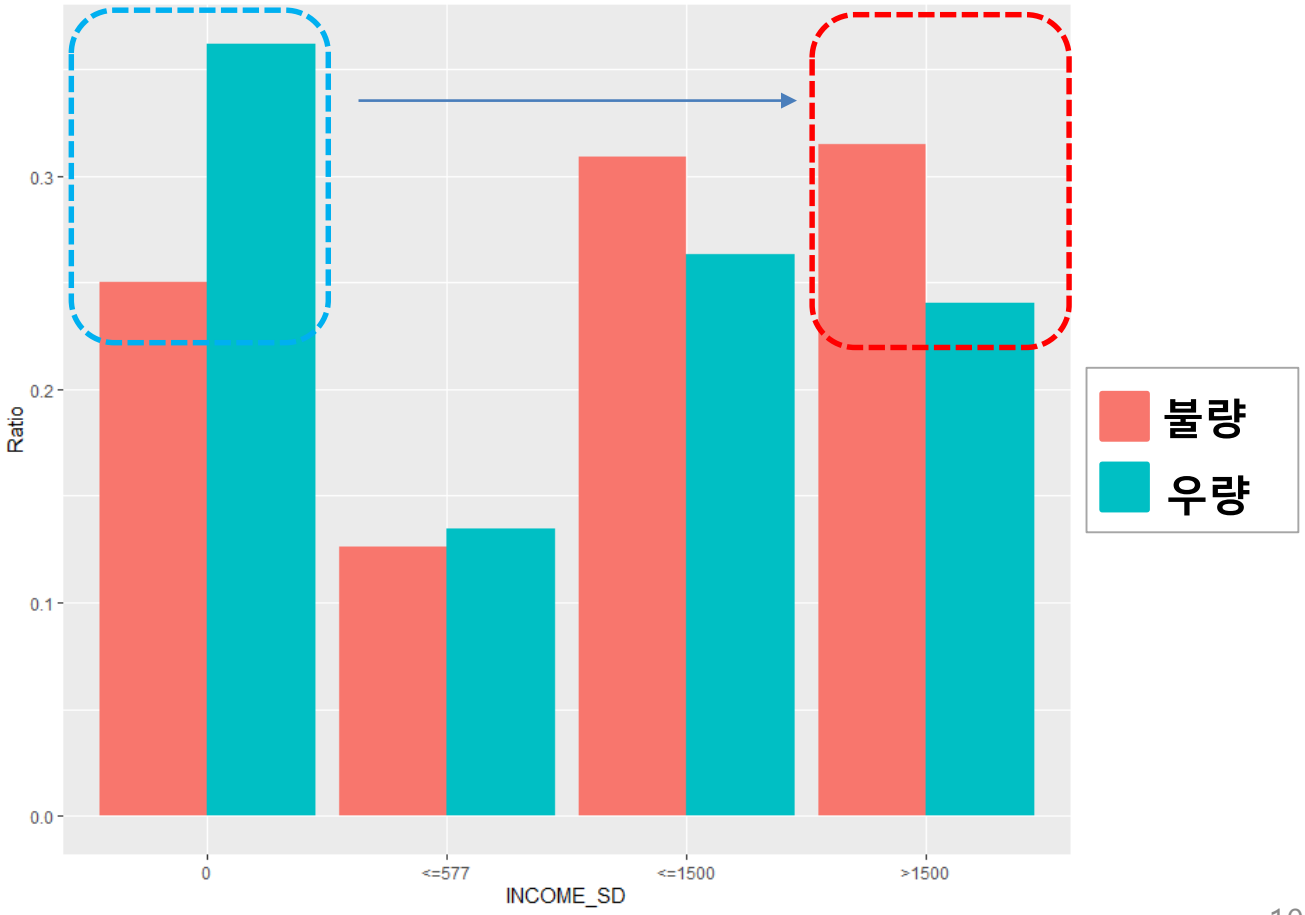
- 연 소득의 변동이 클수록 불량 확률이 증가할 것으로 추측

e.g.) 차주ID: 120009530632

BS_YR_MON	INCOME	DLQ_N1YE_FLAG
201609	22,639,720	0
201612	2,253,700	1
201703	90,000	1
201706	829,630	1
201709	1,590,690	1
201712	77,000	0
201803	75,000	0
201806	80,000	0

4. 최근 1년 간 연 소득 표준편차

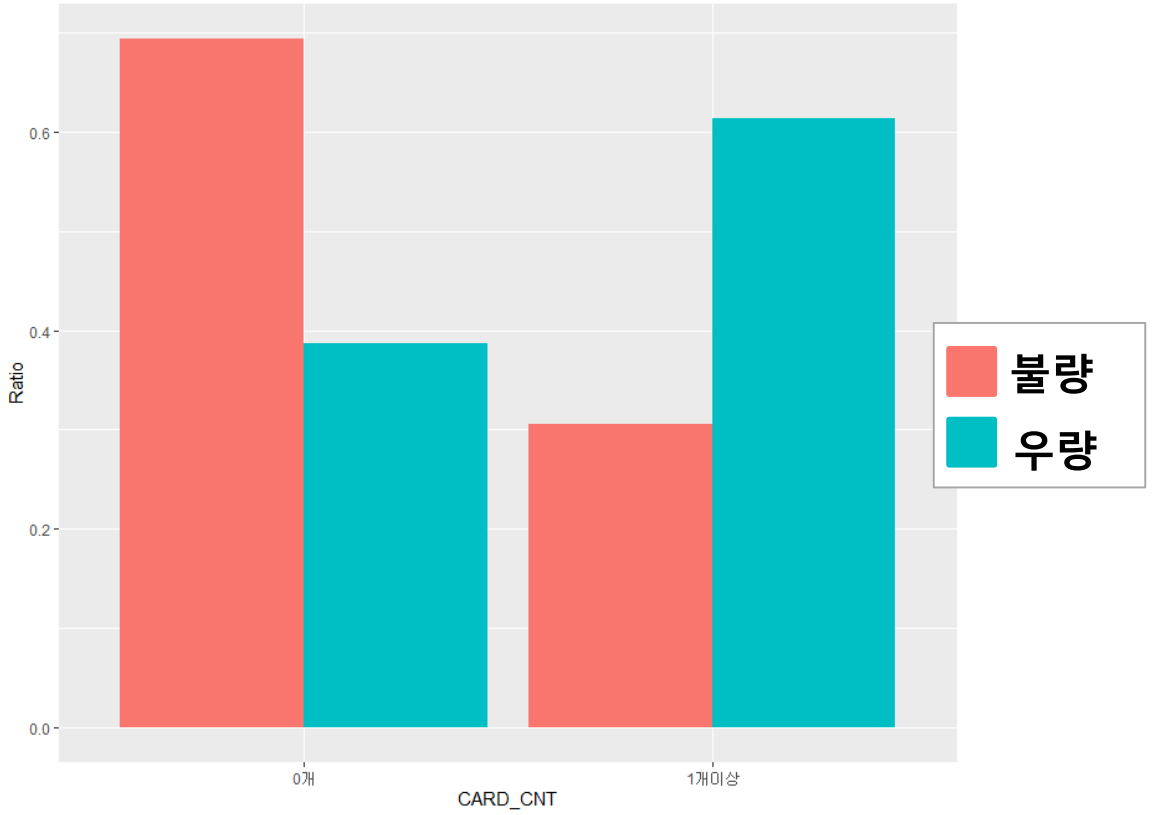
- 확인해본 결과, 연 소득 **표준편차가 클 때 불량률이 높다**



5. 최근 1년 간 신용카드 평균 이용 기관 수 (factor)

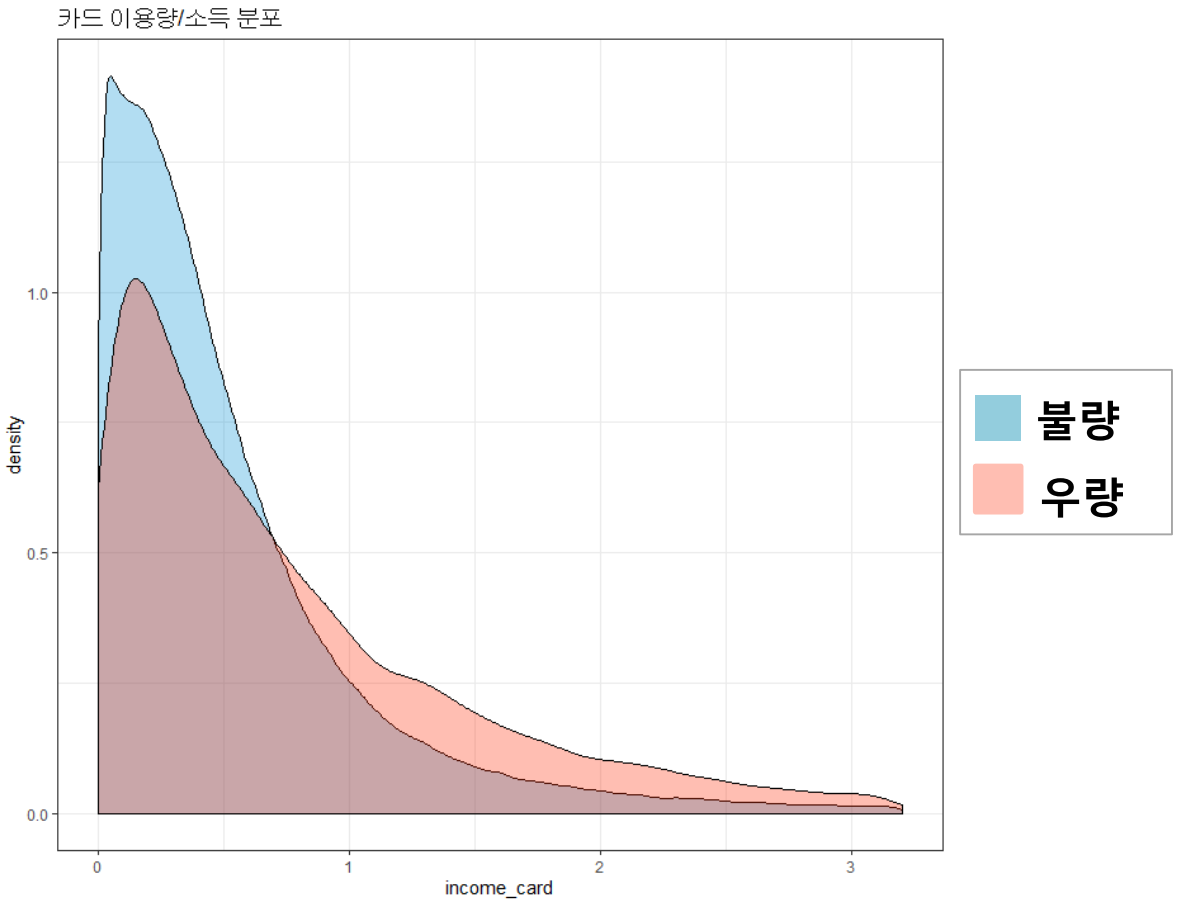
- 이용 기관 수를 0개와 1개 이상으로 구분: **신용카드 이용자가 비이용자에 비해 우량 비율이 높음**

Summary of CARD_CNT	
MIN	0.0
1st Quartile	0.0
Median	1.0
3rd Quartile	1.166
MAX	11.250



6. 최근 1년 간 소득 대비 신용카드 평균 이용금액 (numeric)

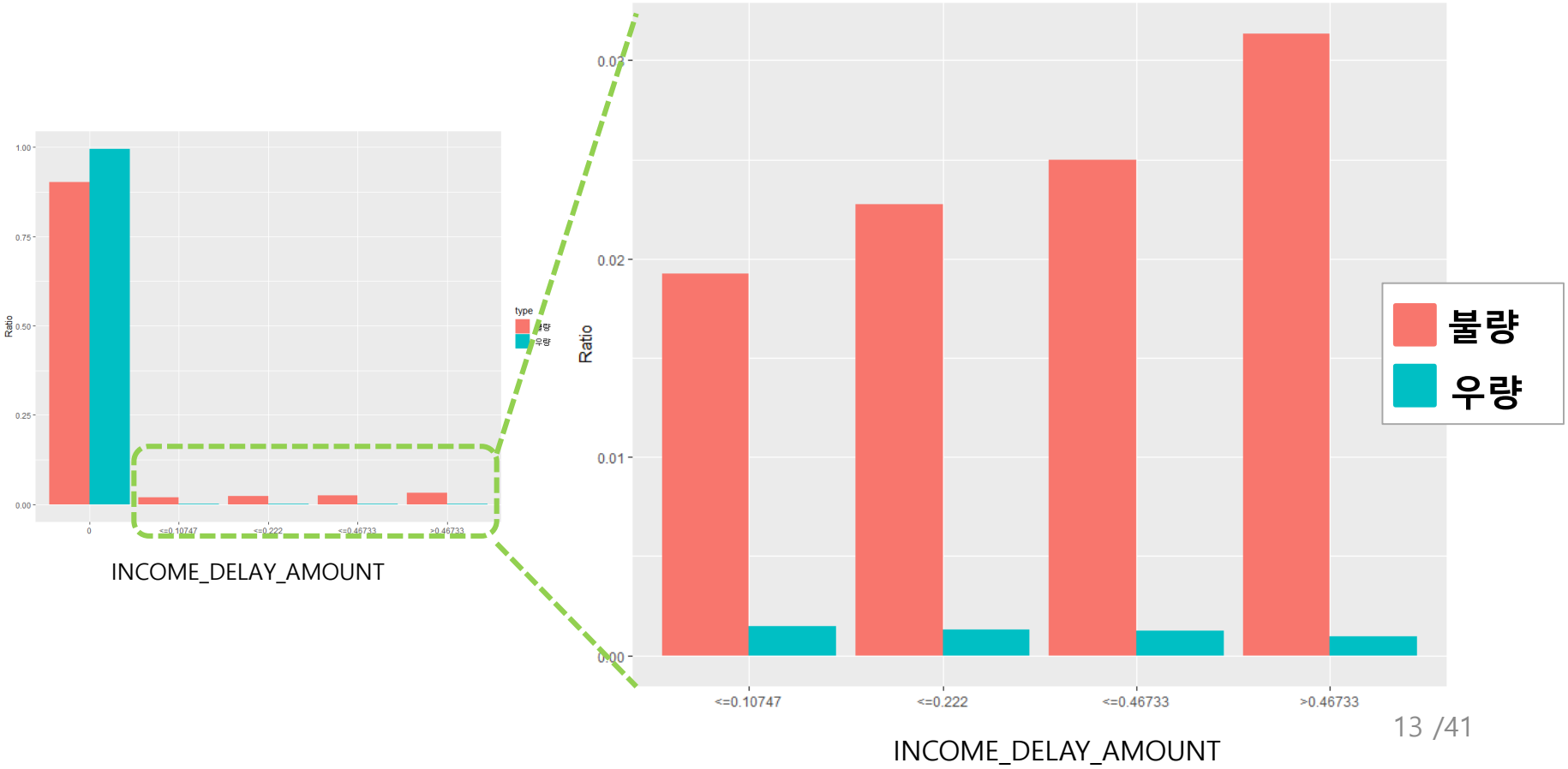
$$\text{INCOME\_CARD}_i := \frac{1}{n}\sum \frac{\text{월별 신용카드 이용금액}}{\text{월별 소득}}$$



7. 최근 1년 간 소득 대비 연체 잔액 평균(number)

$$\text{INCOME\_DELAY\_AMOUNT}_i := \frac{1}{n}\sum \frac{\text{연체잔액}}{\text{연소득}}$$

- 연소득 대비 총 연체잔액이 클수록 불량률이 높음

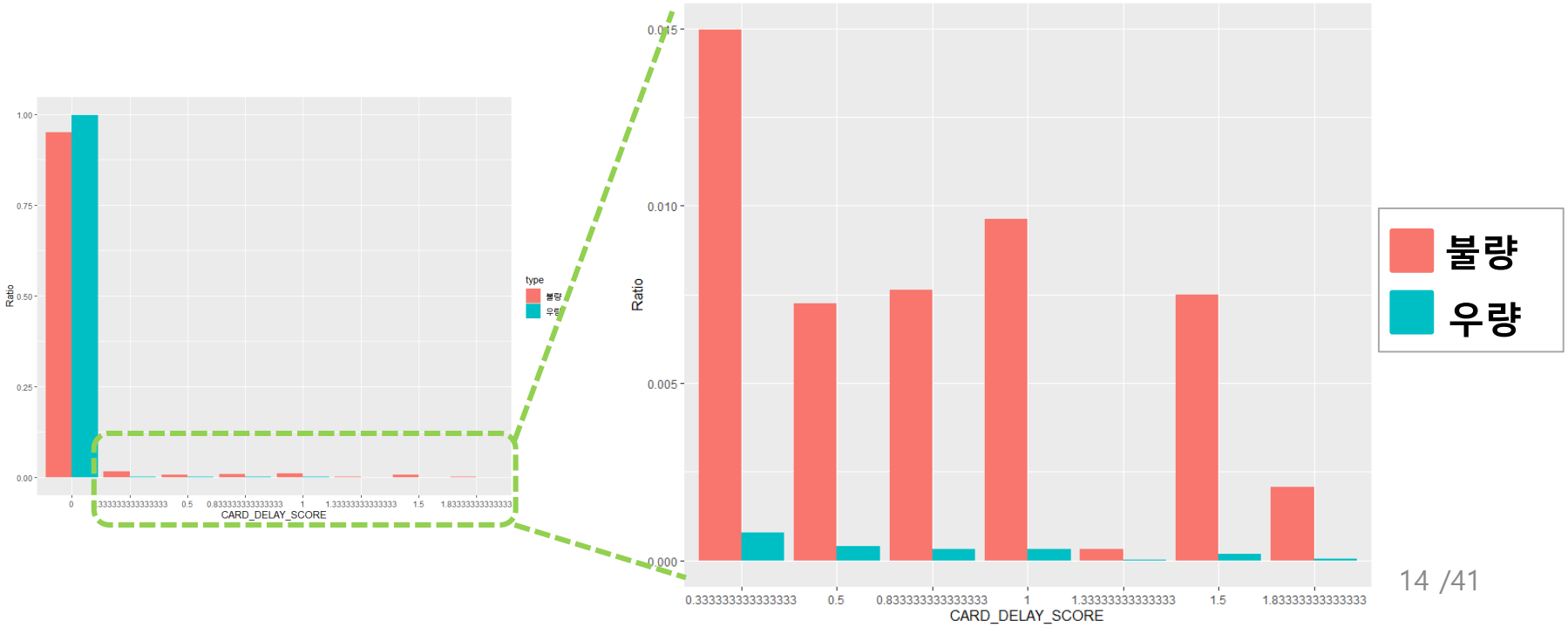


8. 신용카드 연체 점수 (number)

$$\text{CARD\_DELAY\_SCORE}_i := \sum \frac{1}{\text{기준시점 분기} - \text{연체시점 분기}}$$

- 과거 1년 동안의 연체 여부를 이용해 점수 계산
- 연체시기가 기준시점에 가까울수록 높은 가중치를 부여

Ex) 201706 기준 201609 연체, 201703 연체→ CARD_DELAY_SCORE = 1/3 + 1/1 = 1.3333333

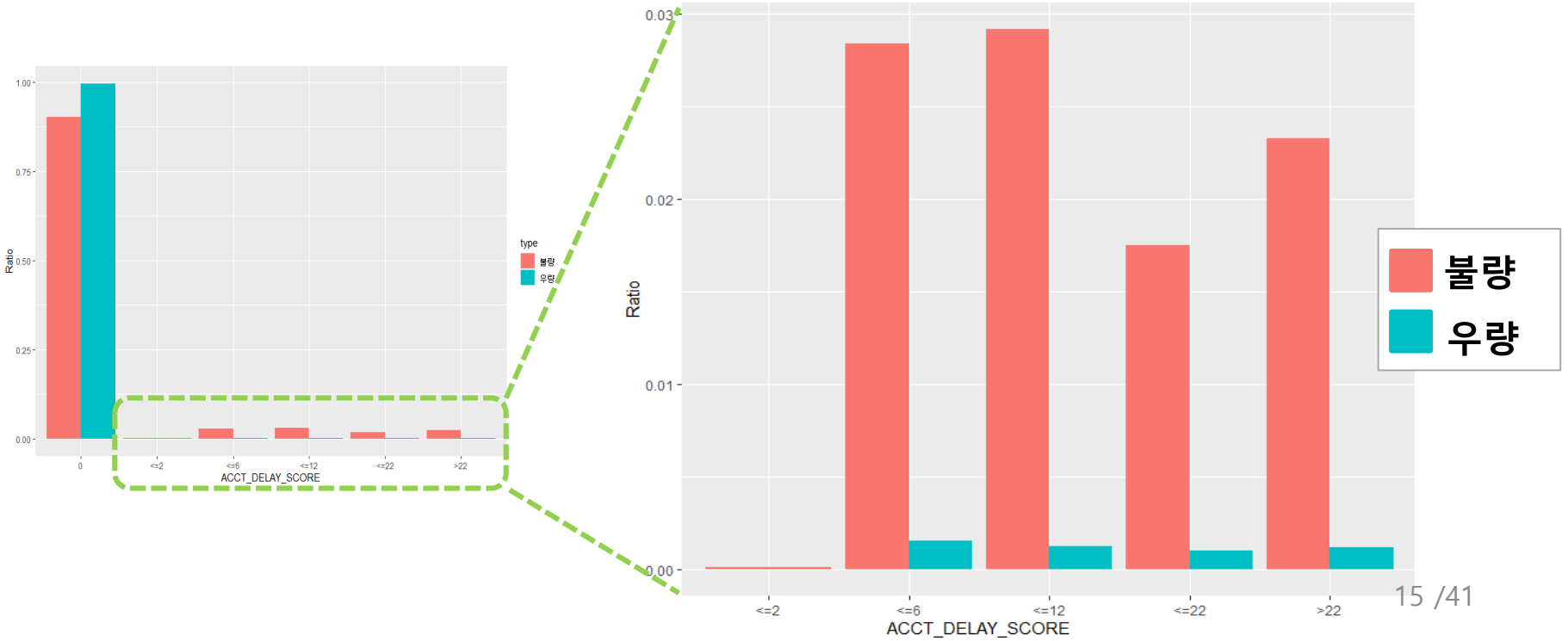


9. 대출 연체 점수 (number)

$$\text{ACCT\_DELAY\_SCORE}_i := \sum \frac{1}{\text{기준시점 분기} - \text{연체시점 분기}}$$

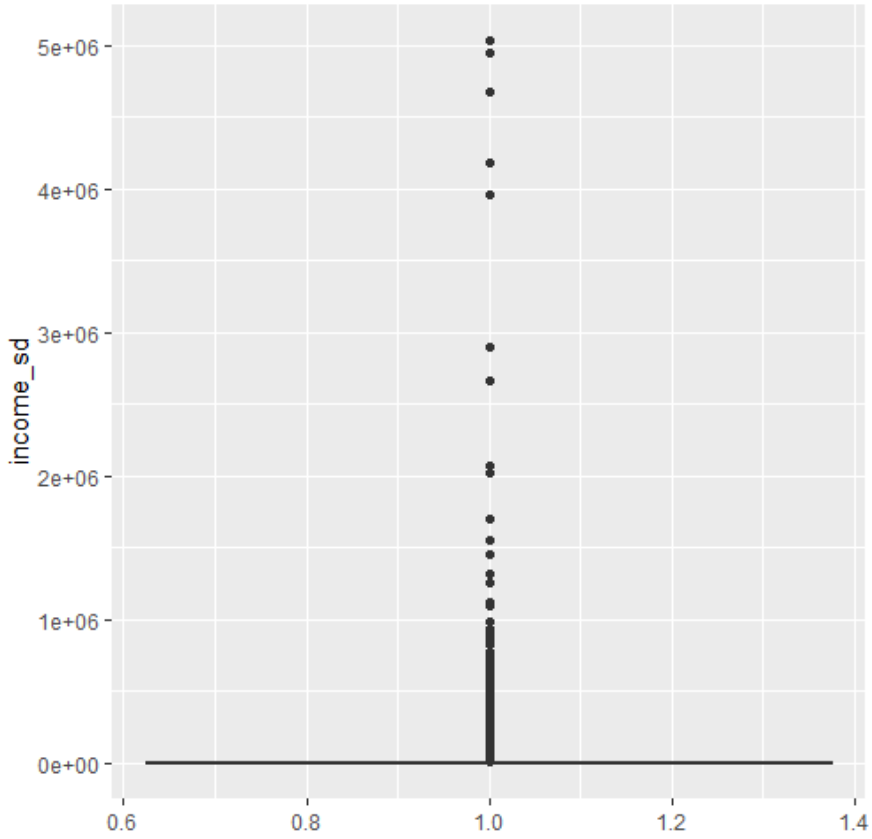
- 과거 1년 동안의 연체 여부를 이용해 점수 계산
- 연체시기가 기준시점에 가까울수록 높은 가중치를 부여

Ex) 201706 기준 201703 연체 + 201606 연체 → ACCT_DELAY_SCORE = 1/3 + 1/1 = 1.3333333



Robust Normalization

- INCOME_SD, DTI 등 변수들의 경우, 극단적으로 높은 값들이 존재.
- 극단적인 값에 Robust할 수 있도록, Robust Normalization 방법을 사용.



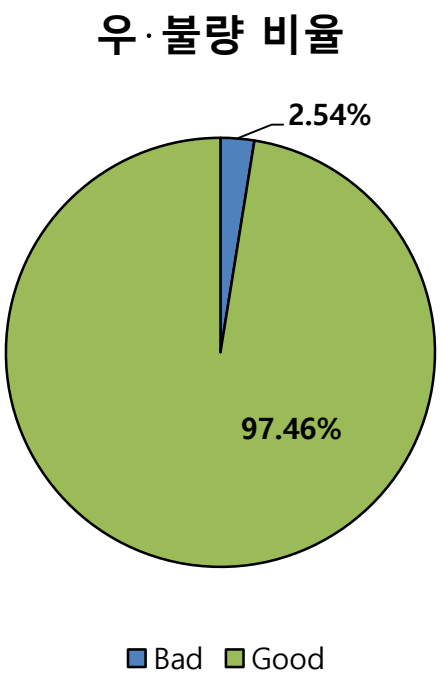
<Outlier 예시>

Robust Normalization

$$x_{norm,\,i} = \sum \frac{x_i - q_2}{q_3 - q_1}$$

- q_i: i 번째 quartile.

Ⅱ-(1). 샘플링 – 클래스 불균형 문제



불량(타겟 클래스) 비율: 2.54%

=> 심각한 클래스 불균형

- **문제점**

- 다수 클래스(우량)을 잘못 예측하는 것보다 소수 클래스(불량)을 잘못 예측하는 것의 비용이 더 큼

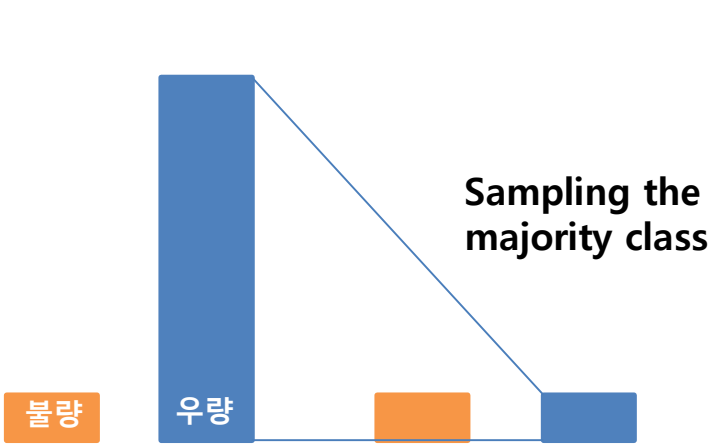
- **대처 방법**

- Undersampling, Undersampling + Ensemble, Balance cascade, Oversampling 등

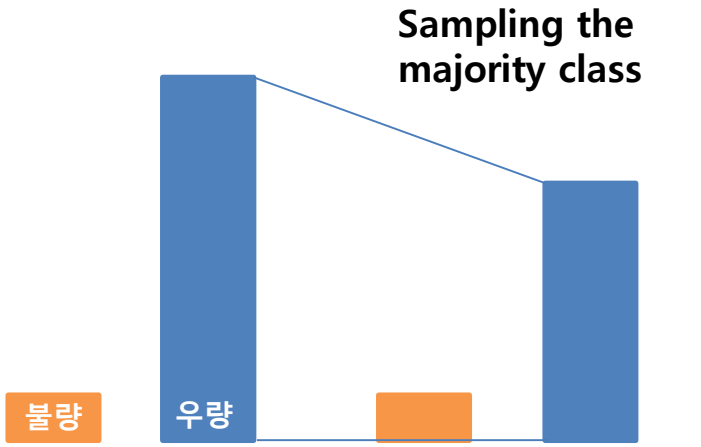
1. Undersampling

- 다수 클래스의 데이터를 랜덤 샘플링

- 1:1, 1:9 비율로 undersampling

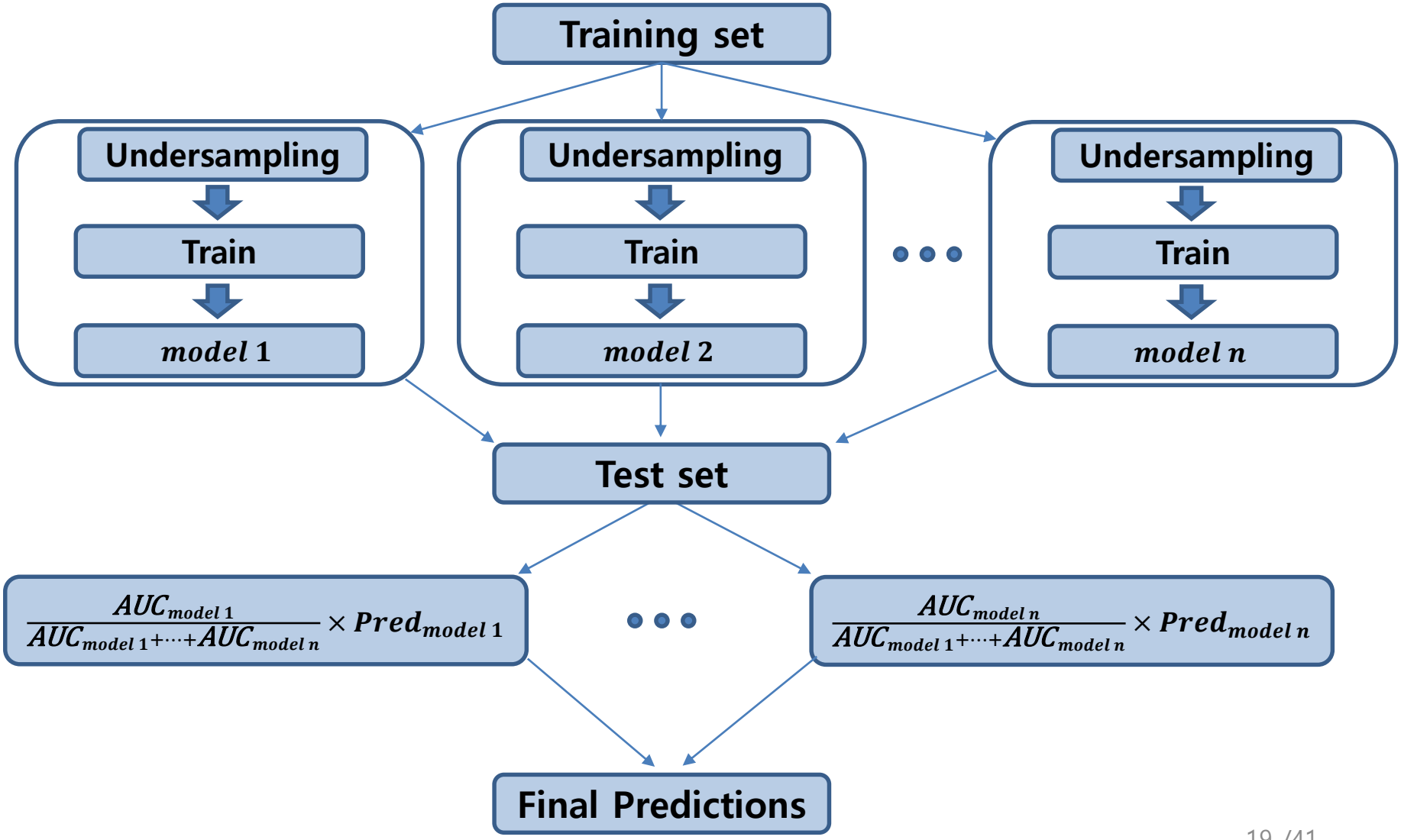


Original dataset 1:1 Undersampling

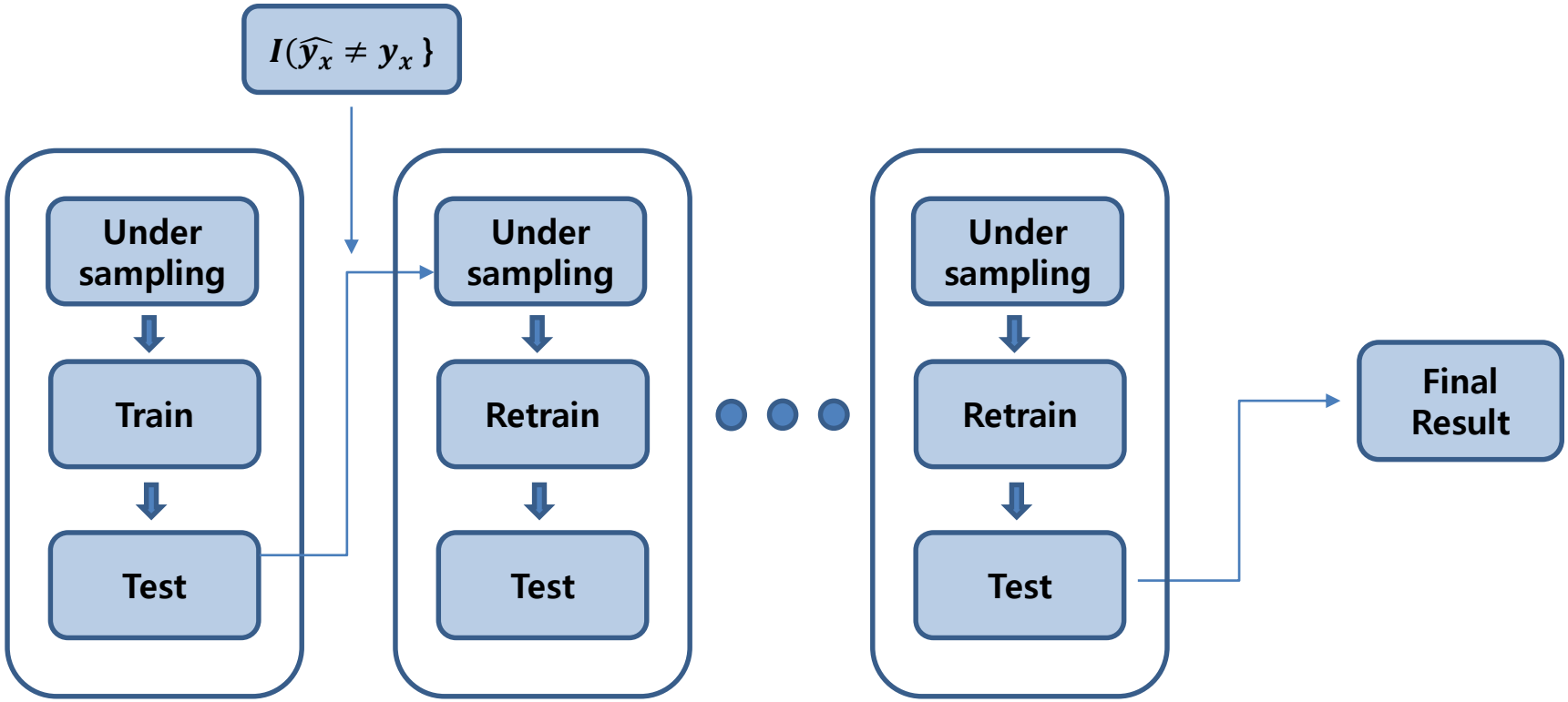


Original dataset 1:9 Undersampling

2. Undersampling + Ensemble

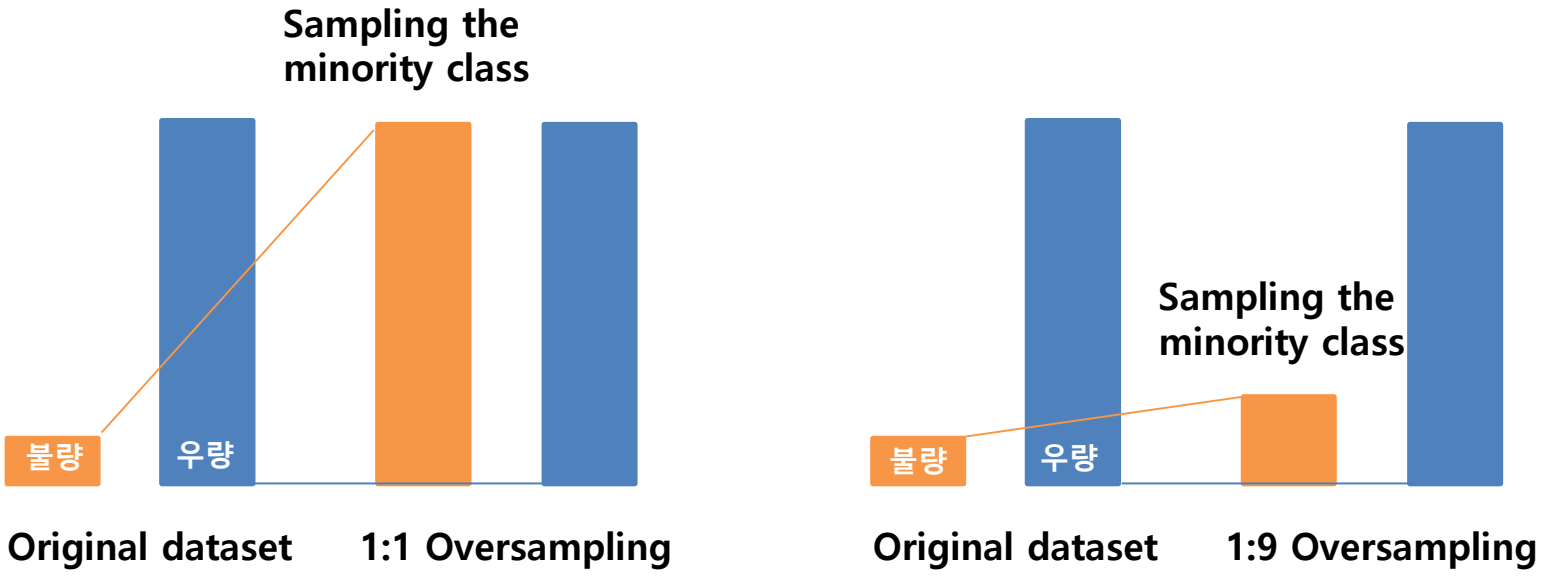


3. Balance cascade

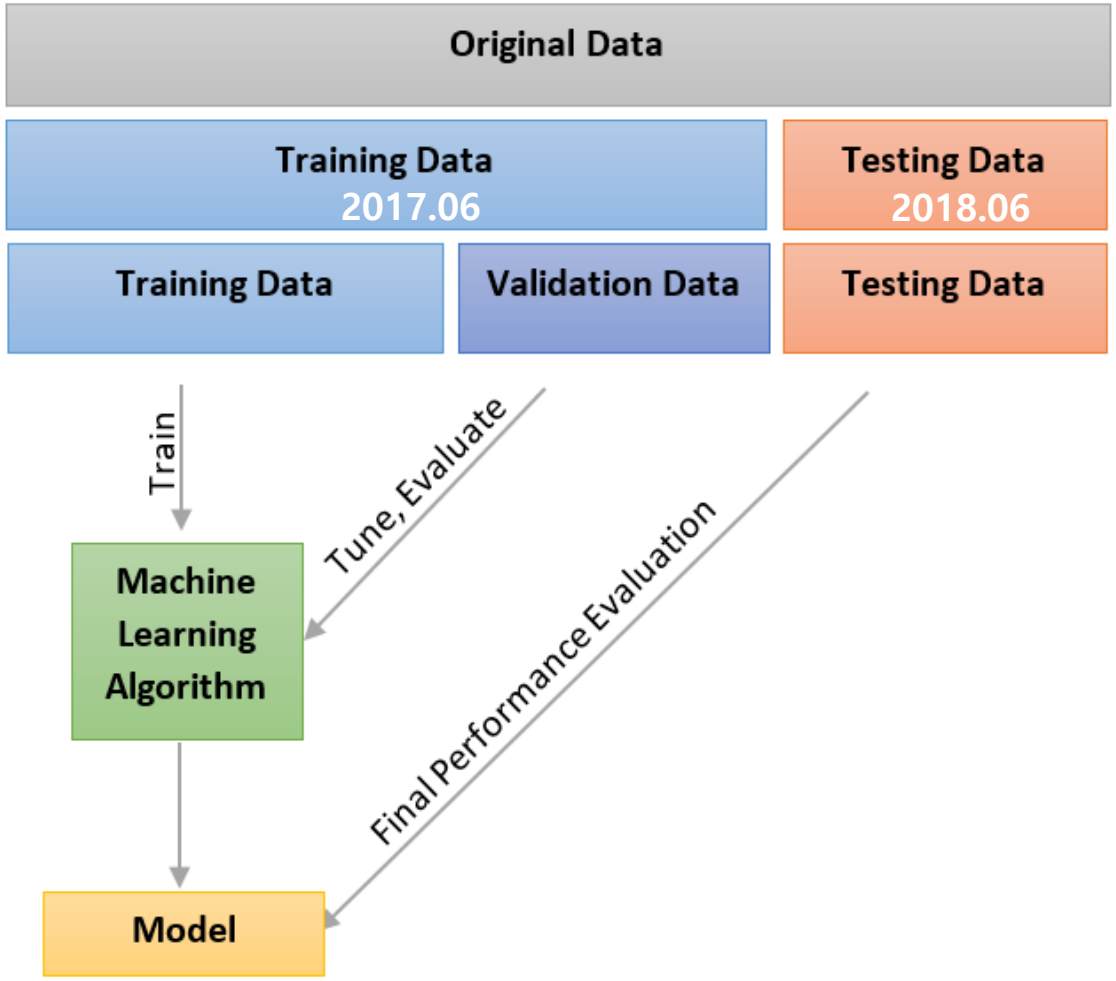


4. Oversampling

- Undersampling은 데이터의 일부를 버리는 것이므로 정보의 손실 발생
- 소수 클래스의 데이터를 반복 추출
- 1:1, 1:9 비율로 oversampling



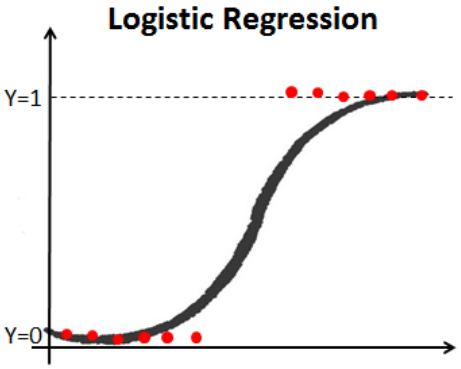
Ⅱ-(2). 모델링 – 데이터 분할


Original Data
Training Data
2017.06
Testing Data
2018.06
Training Data
Validation Data
Testing Data
Train
Tune, Evaluate
Final Performance Evaluation
Machine Learning Algorithm
Model

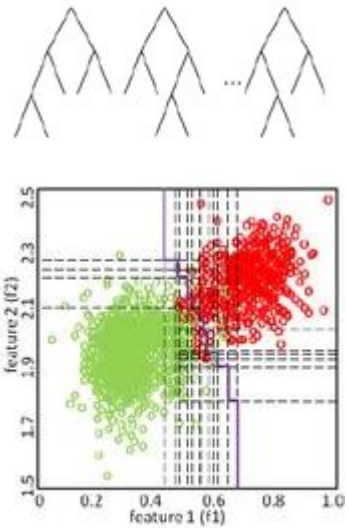
Ⅱ-(2). 모델링 – 사용 모형

로지스틱 회귀

$$\log\frac{p}{1-p} = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

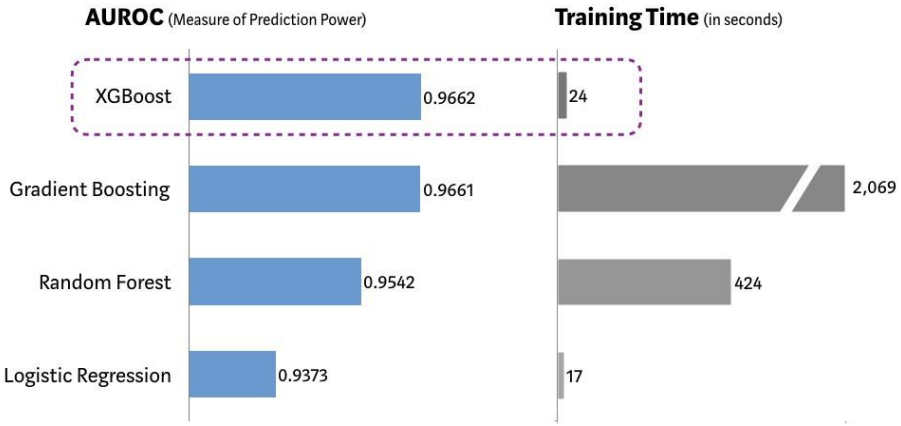


랜덤 포레스트



XGBoost

Performance Comparison using SKLearn's 'Make_Classification' Dataset
(5 Fold Cross Validation, 1MM randomly generated data sample, 20 features)



Ⅱ-(3). 모형 변별력 확인

Positive class = 불량

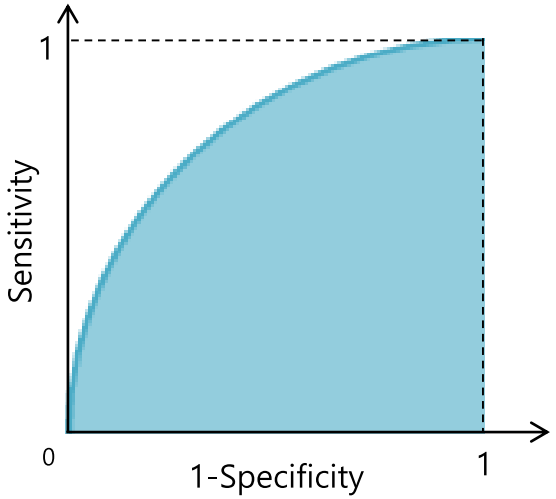
	실제 불량	실제 우량
예측 불량	TP	FP
예측 우량	FN	TN

$$Sensitivity = Recall = \frac{TP}{TP + FN}$$

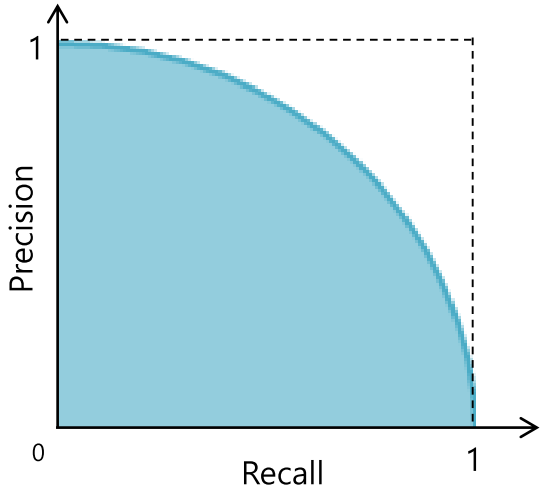
$$Specificity = \frac{TN}{TN + FP}$$

$$Precision = \frac{TP}{TP + FP}$$

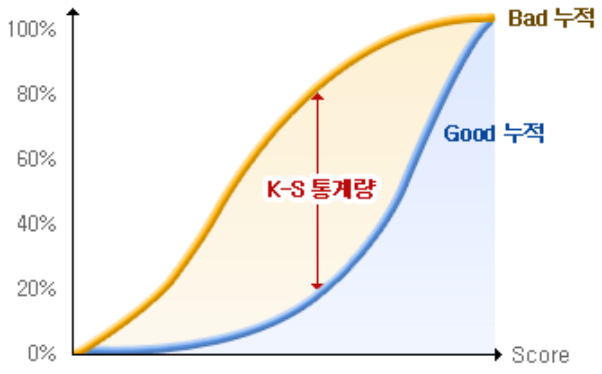
- **ROC-AUC**: ROC 곡선의 아래 면적



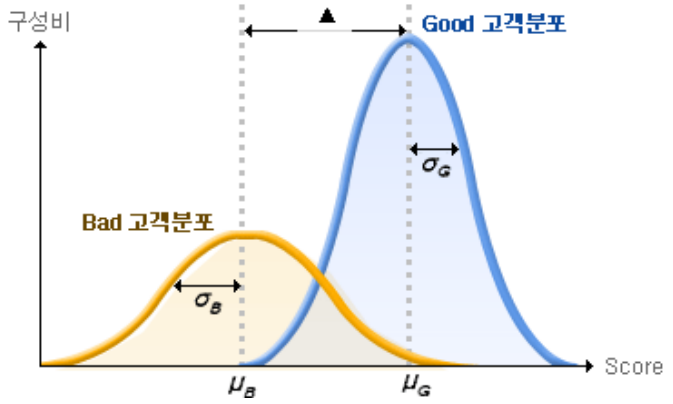
- **PR-AUC**: Precision-Recall 곡선의 아래 면적



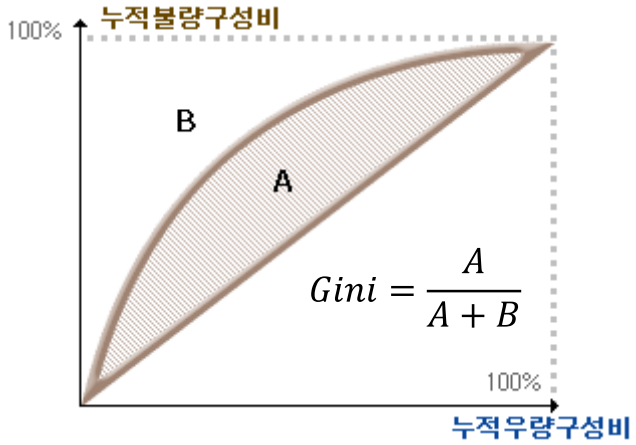
- **K-S Statistics**



- **Divergence**



- **GINI index**



성능지표	적정 기준치
K-S(Kolmogorov-Smirnov)	50 이상
Divergence	1 이상
Gini	0.6 이상

[출처] 올크레딧

Ⅱ-(3). 모형 변별력 확인

Baseline (Validation)				
	AUC	PRAUC	KS	GINI
LR	0.765	0.146	38.24	0.530
RF	0.797	0.176	44.53	0.593
XGB	0.829	0.206	50.38	0.658

Undersampling 1:1 (Validation)				
	AUC	PRAUC	KS	GINI
LR	0.806	0.185	46.80	0.613
RF	0.841	0.198	52.30	0.682
XGB	0.846	0.242	52.70	0.692

Undersampling 1:9 Ensemble (Validation)				
	AUC	PRAUC	KS	GINI
LR	0.800	0.192	45.53	0.599
RF	0.845	0.255	52.94	0.691
XGB	**0.849**	**0.261**	**52.92**	**0.698**

Oversampling 1:1 (Validation)				
	AUC	PRAUC	KS	GINI
XGB	0.824	0.198	49.44	0.641

Without Sampling (Validation)				
	AUC	PRAUC	KS	GINI
LR	0.794	0.188	44.29	0.587
RF	0.833	0.251	50.90	0.665
XGB	**0.849**	**0.259**	**53.04**	**0.697**

Undersampling 1:9 (Validation)				
	AUC	PRAUC	KS	GINI
LR	0.799	0.189	45.31	0.600
RF	0.844	0.254	52.87	0.688
XGB	**0.849**	**0.257**	**53.01**	**0.698**

Undersampling 1:9 Balance Cascade (Validation)				
	AUC	PRAUC	KS	GINI
LR	0.758	0.187	37.54	0.516
RF	0.805	0.239	46.06	0.611
XGB	0.822	0.255	48.43	0.645

Oversampling 1:9 (Validation)				
	AUC	PRAUC	KS	GINI
XGB	0.824	0.210	49.42	0.640

Ⅱ-(3). 모형 변별력 확인

〈표 Ⅲ-3〉 국내외의 차별방지 입법례

국 가	관련법규	차별금지 항목
한 국	국가인권위법	성별, 종교, 장애, 나이, 사회적 신분, 출신 지역(출생지, 등록기준지, 성년이 되기 전의 주된 거주지 등), 출신 국가, 출신 민족, 용모 등 신체조건, 기혼·미혼·별거·이혼·사별·재혼·사실혼 등 혼인 여부, 임신 또는 출산, 가족 형태 또는 가족 상황, 인종, 피부색, 사상 또는 정치적 의견, 형의 효력이 실효된 전과(前科), 성적(性的) 지향, 학력, 병력(病歷) 등

<KIF 한국금융연구원 금융동향: 분석과 전망 2012년 10월호 제 3부 주요 금융 이슈>

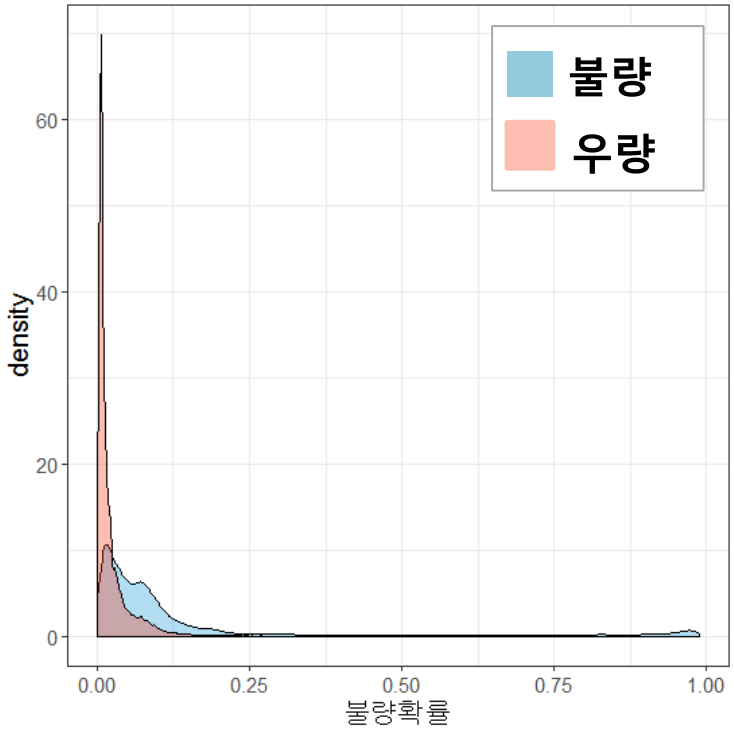
< 민감정보(성별, 나이, 거주지) 제외한 1:9 undersampling 모형>

Undersampling 1:9 (Val)				
	AUC	PRAUC	KS	GINI
LR	0.799	0.189	45.31	0.600
RF	0.844	0.254	52.87	0.688
XGB	**0.849**	**0.257**	**53.01**	**0.698**

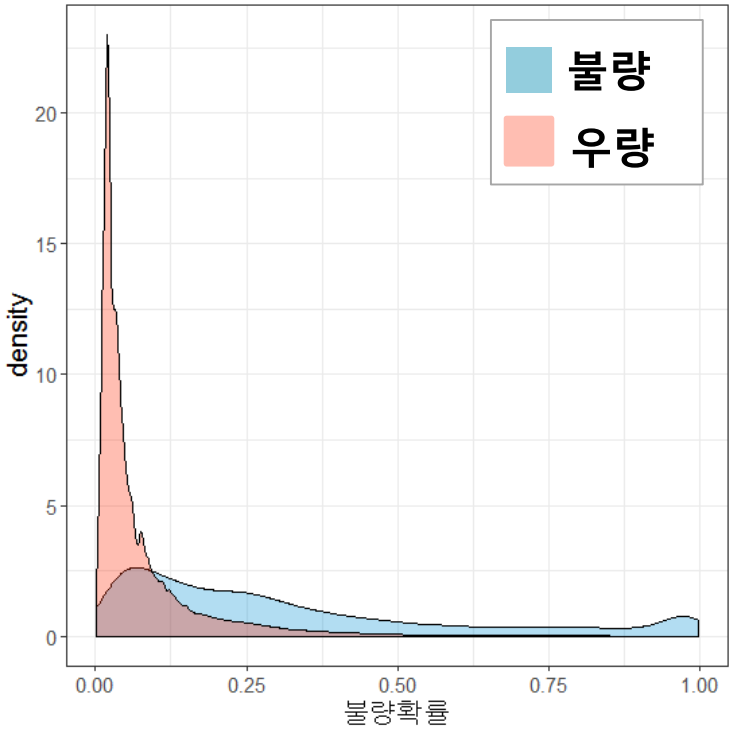
Undersampling 1:9 (Val) - 민감정보 제외				
	AUC	PRAUC	KS	GINI
LR	0.776	0.184	43.38	0.553
RF	0.828	0.247	49.99	0.657
XGB	**0.832**	**0.250**	**50.54**	**0.665**

Ⅱ-(3). 모형 변별력 확인

< Baseline 모형 >



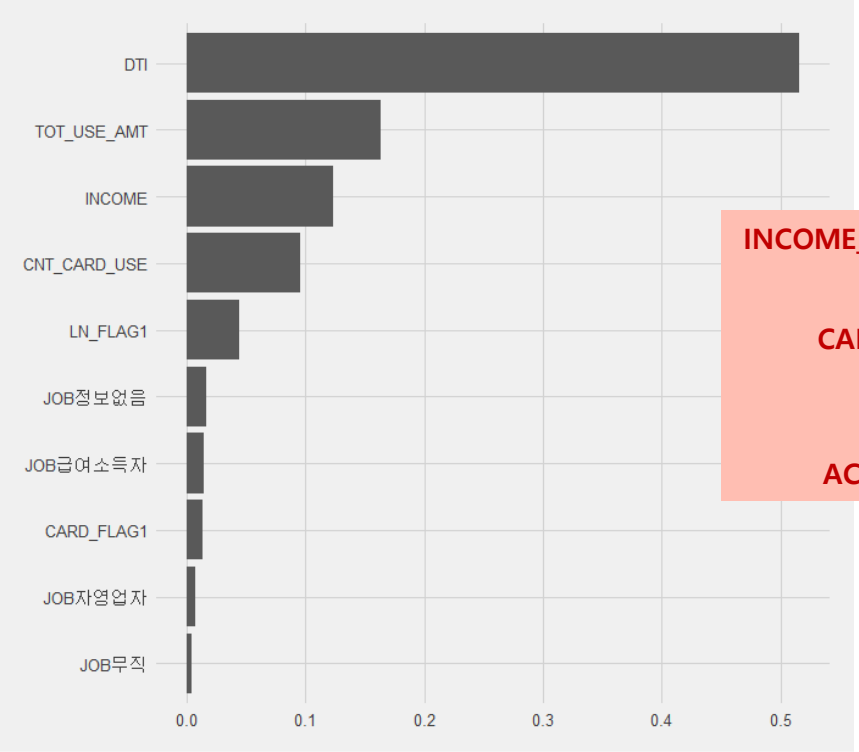
< 최종 모형 >



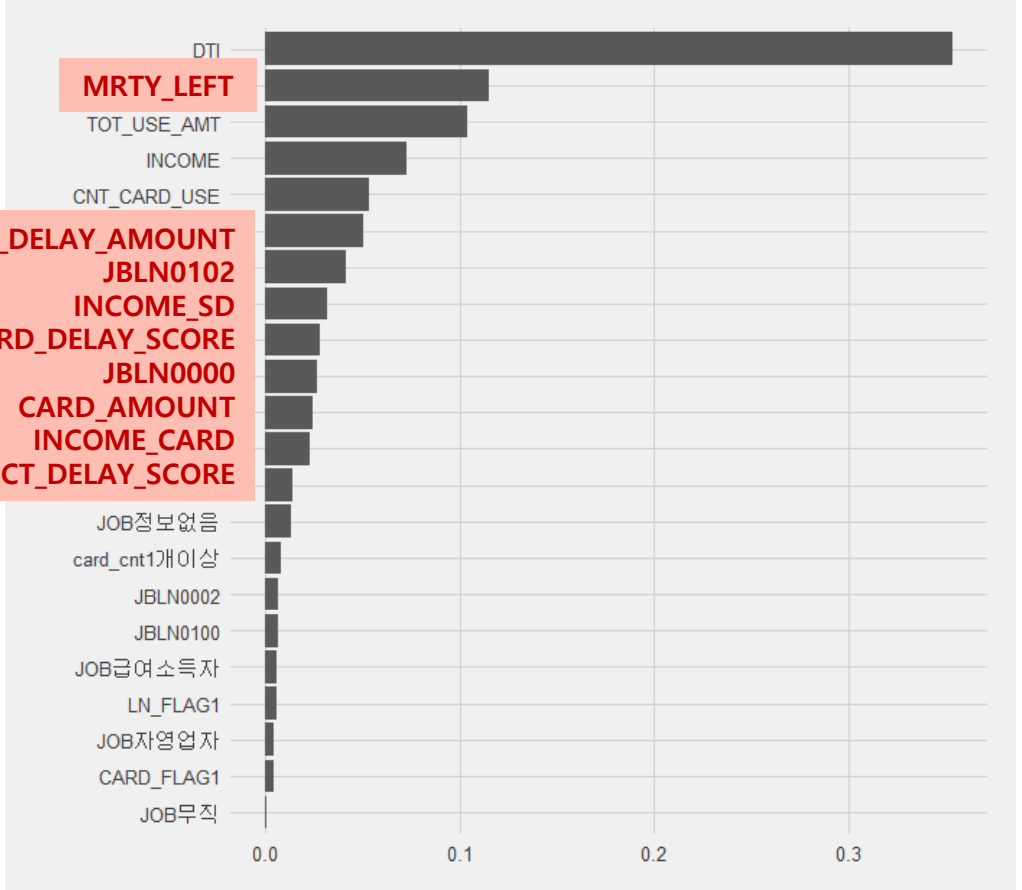
	Baseline (Test)	최종모형 (Test)
KS	50.5396	**53.6529**
Divergence	0.47	**1.36**

Ⅱ-(3). 모형 변별력 확인 – 변수 중요도

<Baseline 모형>

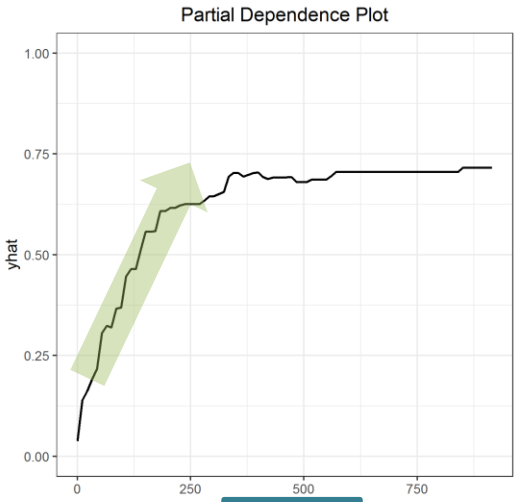


<최종 모형>

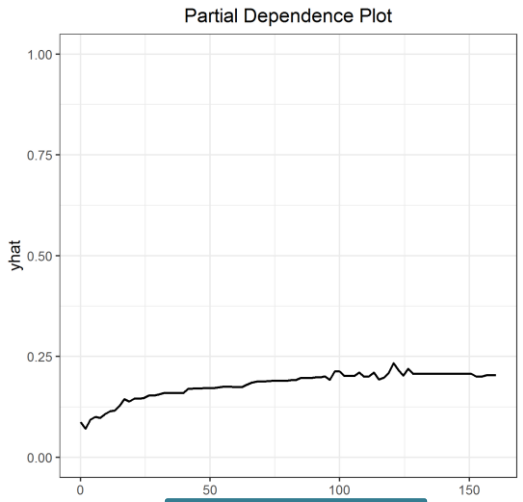


- 추가한 파생변수들의 변수 중요도가 대부분 높게 나타남

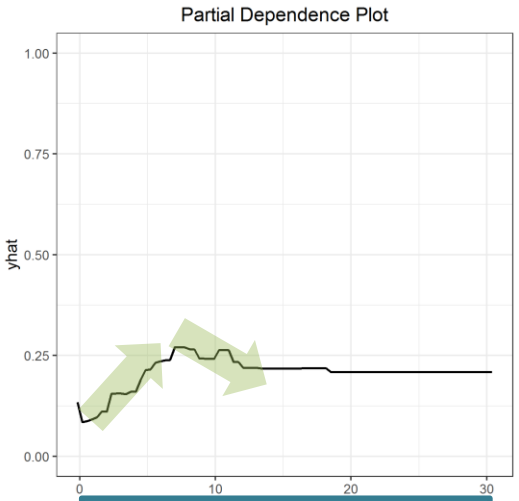
Ⅱ-(3). 모형 변별력 확인 – Partial Dependence Plot



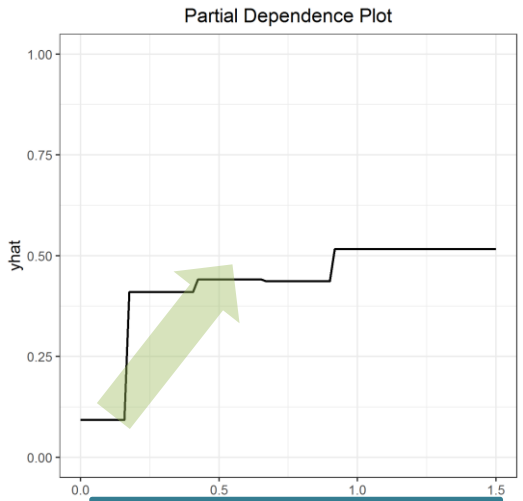
DTI



MRTY_LEFT



TOT_USE_AMT



CARD_DELAY_SCORE

- 모형의 **평균적인 예측결과** 파악
- 변수의 비선형적 패턴, 특히 예측값이 **증가하다 감소하는 패턴**도 파악 가능
- 각 변수가 우·불량 여부에 **어떤 방향**으로 영향을 미치는지 파악 가능

- **PSI(Population Stability Index)**

$$PSI = D_{KL}(\hat{q}(x)|\hat{p}(x)) + D_{KL}(\hat{p}(x)|\hat{q}(x))$$

$$where\ D_{KL}(q(x)|p(x)) = E_p\left(\ln\left(\frac{p(X)}{q(X)}\right)\right) = \sum_{i=1}^{B} p(x_i) ln \frac{p(x_i)}{q(x_i)}$$

$\hat{p}(x)$: Validation set에서 각 cutoff 구간에 속하는 비율

$\hat{q}(x)$: Test set에서 각 cutoff구간에 속하는 비율

[출처] Yurdakul, Bilal, "Statistical Properties of Population Stability Index" (2018). Dissertations. 3208.

⇒ Validation set과 Test set에서 **두 확률분포의 차이**를 계산

	Baseline	최종모델
PSI	0.000505 **<<<<0.1**	0.000622 **<<<<0.1**

(1) 대출상품에 따른 모델 변별력

- 신용대출 보유자와 담보대출 보유자 모두 Baseline 모형보다 **최종모형의 성능이 더 좋았음**

Baseline (Test) - 민감정보 제외				
	AUC	PRAUC	KS	GINI
신용대출	0.7786	0.1999	**40.7136**	0.5572
담보대출	0.7715	0.1453	**40.4249**	0.543

Undersampling 1:9 (Test) - 민감정보 제외				
	AUC	PRAUC	KS	GINI
신용대출	0.8193	0.2617	**47.5050**	0.6385
담보대출	0.8302	0.2423	**50.5845**	0.6605

(2) 업종에 따른 모델 변별력

- 은행권의 경우에는 결과가 매우 좋은 편

- **캐피탈, 저축은행**의 경우에 은행권에 비해 모형의 변별력이 낮음

Baseline (Test) - 민감정보 제외				
	AUC	PRAUC	KS	GINI
은행	0.7881	0.1194	**43.8720**	0.5762
저축은행 +캐피탈	0.7068	0.2397	**28.7530**	0.4136

Undersampling 1:9 (Test) - 민감정보 제외				
	AUC	PRAUC	KS	GINI
은행	0.8423	0.2009	**53.2428**	0.6846
저축은행 +캐피탈	0.7581	0.2986	**36.8929**	0.5162

(2) 업종에 따른 모델 변별력

- 개선 방안

INCOME_DELAY_AMOUNT 변수를 계산할 때 **업종별 불량률**을 곱해서 같은 연체 금액이더라도 **캐피탈, 저축은행의 연체금액에 대한 높은 가중치**를 줌

$$\text{INCOME\_DELAY\_AMOUNT}_i := \sum \left(\frac{\text{업종별 연체잔액}}{\text{연소득}} \times \text{업종별 불량률}\right)$$

(2) 업종에 따른 모델 변별력

- 파생변수(INCOME_DELAY_AMOUNT) 수정한 결과 (전체 차주)

Undersampling 1:9 (Test)				
	AUC	PRAUC	KS	GINI
XGB	0.848	0.267	**53.65**	0.696

Undersampling 1:9 (Test) – weighted mean 적용				
	AUC	PRAUC	KS	GINI
XGB	0.826	0.222	**50.29**	0.650

- 파생변수(INCOME_DELAY_AMOUNT) 수정한 결과 (저축은행 + 캐피탈)

Undersampling 1:9 (Test)				
	AUC	PRAUC	KS	GINI
XGB	0.7581	0.2986	**36.8929**	0.5162

Undersampling 1:9 (Test) – 저축은행+캐피탈				
	AUC	PRAUC	KS	GINI
XGB	0.7653	0.2401	**39.3614**	0.5259

- 전체 차주에 대한 모형의 변별력은 나빠지지만 **저축은행과 캐피탈 업종에 대한 변별력은 좋아짐**

- 분석 의의 ①

- 최종 모형인 1:9 Undersampling 은 다른 샘플링 기법에 비해 **학습에 소요되는 시간이 적으면서 성능은 좋게** 나타남

Undersampling 1:9 (Validation)				
	AUC	PRAUC	KS	GINI
LR	0.799	0.189	45.31	0.600
RF	0.844	0.254	52.87	0.688
XGB	**0.849**	**0.257**	**53.01**	**0.698**

Undersampling 1:9 Ensemble (Validation)				
	AUC	PRAUC	KS	GINI
LR	0.800	0.192	45.53	0.599
RF	0.845	0.255	52.94	0.691
XGB	0.849	0.261	52.92	0.698

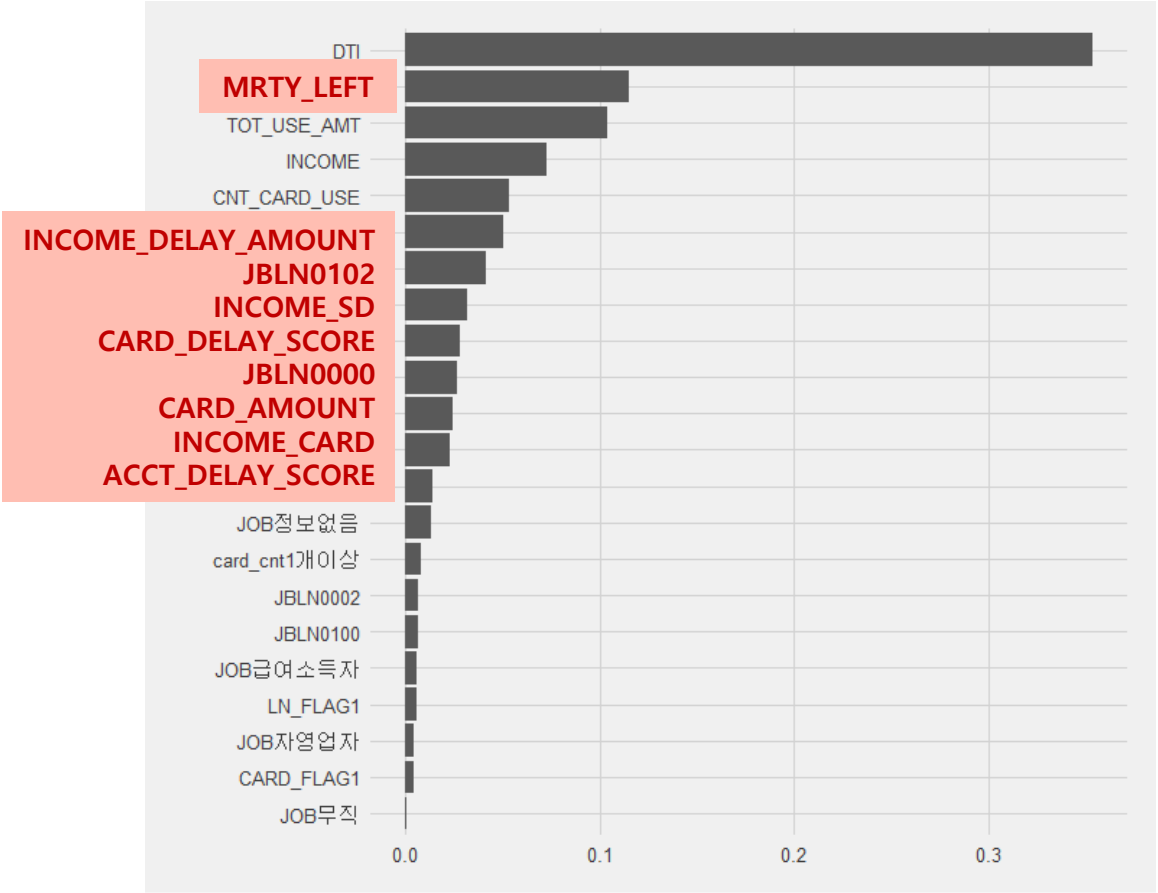
Undersampling 1:9 Balance Cascade (Validation)				
	AUC	PRAUC	KS	GINI
LR	0.758	0.187	37.54	0.516
RF	0.805	0.239	46.06	0.611
XGB	0.822	0.255	48.43	0.645

III-(1). 분석의의

- 분석 의의 ②

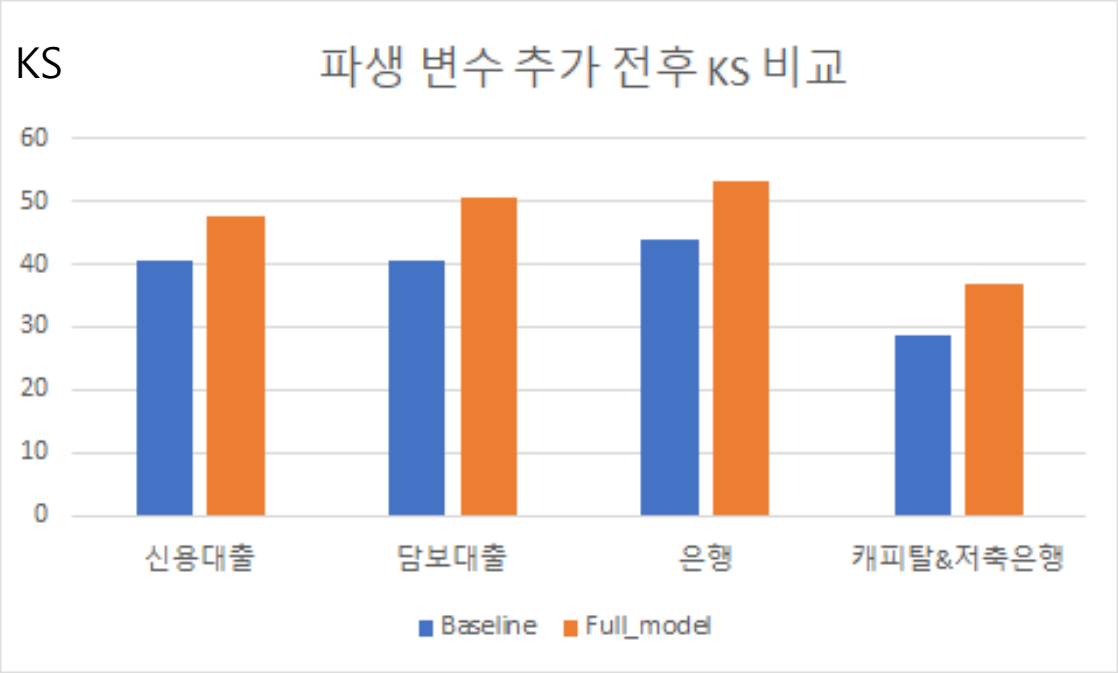
- 추가한 파생변수들의 변수 중요도가 대부분 높게 나타남

변수중요도



- 분석 의의 ③

- 모든 업종 및 상품에 효과적인 파생변수 생성



- 전체 모집단 외 특정 집단 대상으로도 변별력 있도록 모형 구성

III-(2). 발전방향

(1) 대출정보가 없는 차주에 의해 변별력이 낮아지는 현상

- 파생변수는 대부분 대출정보를 활용한 것인데 **대출정보가 없는 사람들은 값이 모두 0으로 들어가는 문제**
- 대출정보 보유 여부로 차주를 나누어 최종모형의 변별력을 확인했을 때, **대출정보가 없는 경우 대출정보가 있는 경우에 비해 변별력이 매우 낮음**
- 대출정보가 있는 차주와 없는 차주 각각에 대해 개별 모형을 만들어 예측한 결과도 크게 다르지 않았음

Undersampling 1:9 (Test) – 전체모형에서 예측				
	AUC	PRAUC	KS	GINI
대출정보 O	0.8304	0.2486	**50.3447**	0.6548
대출정보 X	0.7666	0.1230	**41.3202**	0.5459

Undersampling 1:9 (Test) – 개별 모형에서 예측				
	AUC	PRAUC	KS	GINI
대출정보 O	0.8291	0.2491	**50.2979**	0.6522
대출정보 X	0.7891	0.1239	**42.9453**	0.5718

III-(2). 발전방향

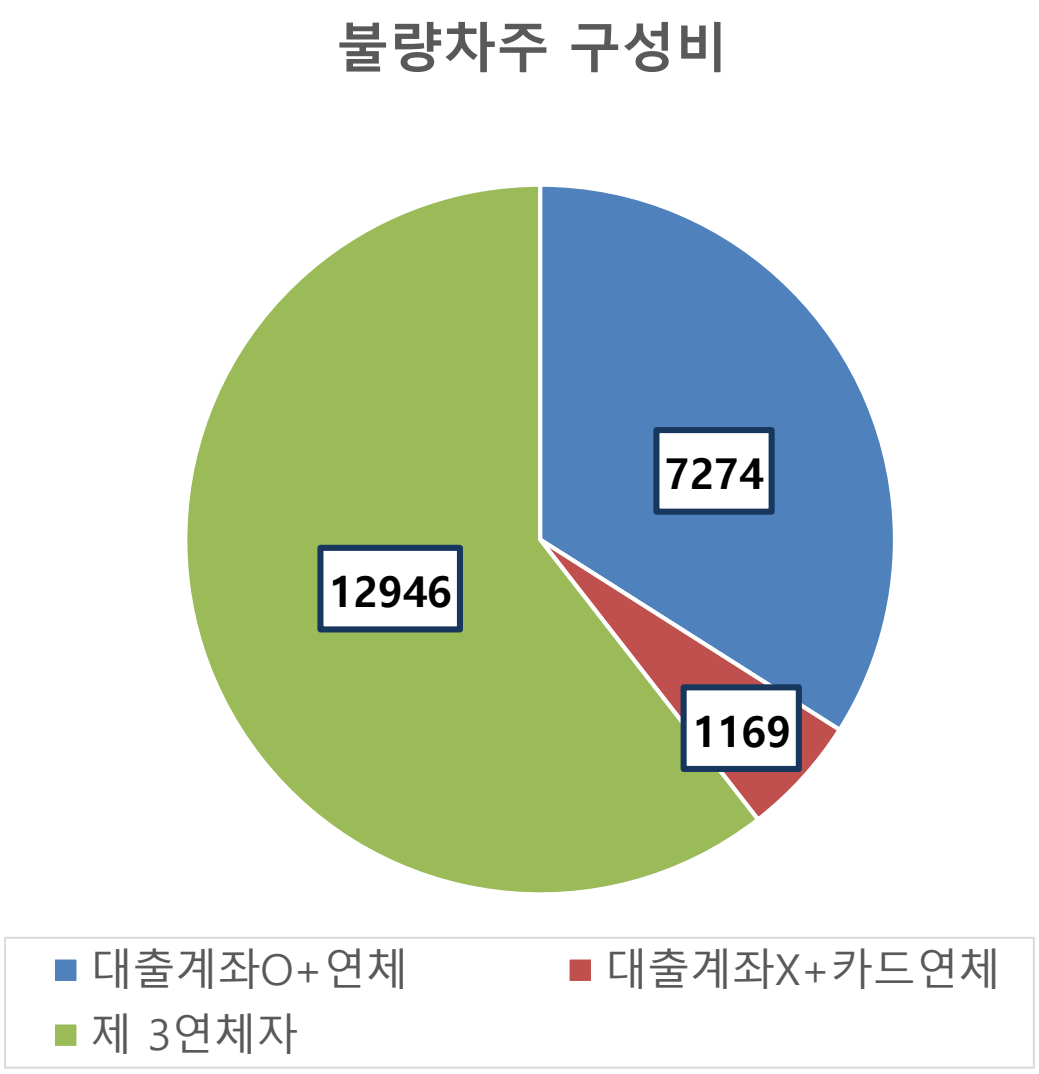
(1) 대출정보가 없는 차주에 의해 변별력이 낮아지는 현상

- 비금융 정보 활용
- 최종모형에서 연체관련 변수의 중요도가 높게 나온 것을 고려해볼 때 통신요금, 공공요금 등의 연체기록을 이용하여 파생변수를 생성하면 모형의 변별력을 높일 수 있을 것으로 기대됨

정보명	활용기간	평가내용	등록방법
소득금액증명	등록일로부터 18개월	신용여력	증명서 발급 후 등록
건강보험	등록일로부터 18개월	신용여력, 신용성향 (약속이행)	증명서 발급 후 등록
국민연금	등록일로부터 18개월	신용여력, 신용성향 (약속이행)	증명서 발급 후 등록
통신요금	등록일로부터 6개월	신용성향 (약속이행)	납부확인서 발급 후 등록
공공요금	등록일로부터 6개월	신용성향 (약속이행)	납부확인서 발급 후 등록
납부내역증명 (납세사실증명)	등록일로부터 18개월	신용성향 (약속이행)	증명서 발급 후 등록

[출처] 올크레딧

(2) 제 3 연체자(대출/카드 연체 기록이 없는 불량 차주)



- 연체관련 정보는 한국신용정보원, 개인 신용정보사의 채무불이행 정보, 기타 연체 공공기록정보와 금융질서문란 정보 등을 포함

[출처] 올크레딧

감사합니다