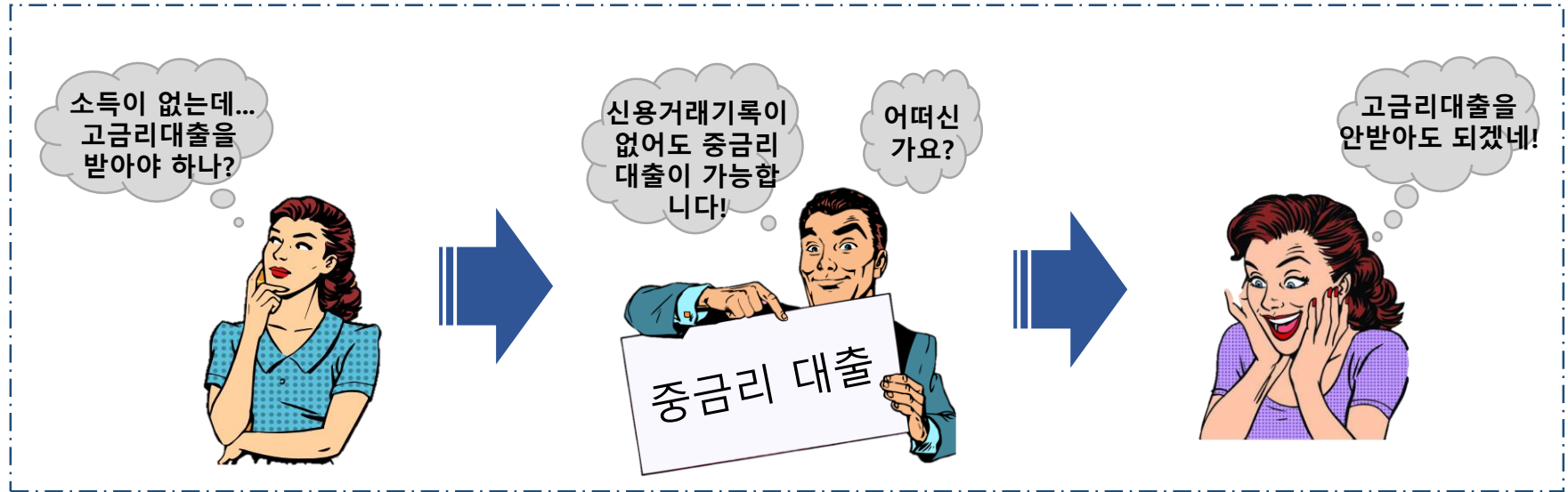


배깅 알고리즘을 적용한 XGboost 대출상환 예측 알고리즘 개발

기조
이정호 김기조 이웅기 김원중 김다영

◆ 중금리 대출



◆ 기존 대출의 한계점

- 중신용자의 고금리대출
 - 주부, 학생 등 중신용자들의 저금리대출이 어려움
 - 상환능력이 있음에도 고금리대출을 받음

중신용자의 데이터로부터 대출상환 예측 알고리즘을 개발하여 중금리대출 활성화

대출상환 예측 알고리즘 개발

- Preprocessing
- Feature value
- Imbalanced data
- Model
- Conclusion

◆ 데이터 설명

- 대출연체여부 예측을 위한 변수 68개
- 한국 감정원 공공 API로 수집한 변수 2개
- 관측치 100,233개

	구분	설명	변수 개수
DATA	SCI	대출 건수 / 대출 금액 / 대출계좌 유지 개월 / 보증 건수 및 금액 등 14개	14개
	한화생명	직업 / 소득 / 가구 정보 / 신용등급 / 한화생명 신용대출정보 / 보험 정보 등 36개	36개
	인구통계학 변수	연령 / 성별	2개
	SKT	통화량 / 멤버십 / 요금 / 회선 상태 / 단말기 등 15개	15개
	한국 감정원	주택매매가격지수 / 전월 대비 변화 율	2개

◆ 변수 변환

- 대출건수(은행,캐피탈,2산업,기타)
 - 각 대출 출처의 가중치를 다르게 고려
 - 대출건수변환공식 => $\text{대출건수} / (\text{해당 변수의 평균 대출 건수}) * \text{대출건수} / \text{해당 사람의 총 대출 건수}$
 - 대출건수/평균 대출 건수 : 일반적인 대출비율
 - 대출 건수/해당 사람의 총 대출 건수 : 해당 사람의 해당 변수에 대한 대출 비율
 - 변수 변환 전 은행 대출 건수에서는 같은 1을 가지는 값이 변환 후 은행 대출 건수에서는 1.19와 0.39와 같이 다른 값으로 나타남
 - > 이는 해당 고객이 은행에서의 대출하는 비율이 높다는 것을 나타냄 (가중치 부여)

은행 대출 건수	카드/할부사/ 캐피탈 대출 건수	2산업 대출 건수	기타 대출 건수
1	0	0	0
1	0	0	0
0	1	3	2
0	2	4	2
4	0	0	0
1	0	1	1



은행 대출 건수	카드/할부사/ 캐피탈 대출 건수	2산업 대출 건수	기타 대출 건수
1.190319314	0	0	0
1.190319314	0	0	0
0	0.332501	1.573511	1.39477
0	0.997503	2.098015	1.046078
4.761277255	0	0	0
0.396773105	0	0.349669	0.697385

◆ 변수 변환

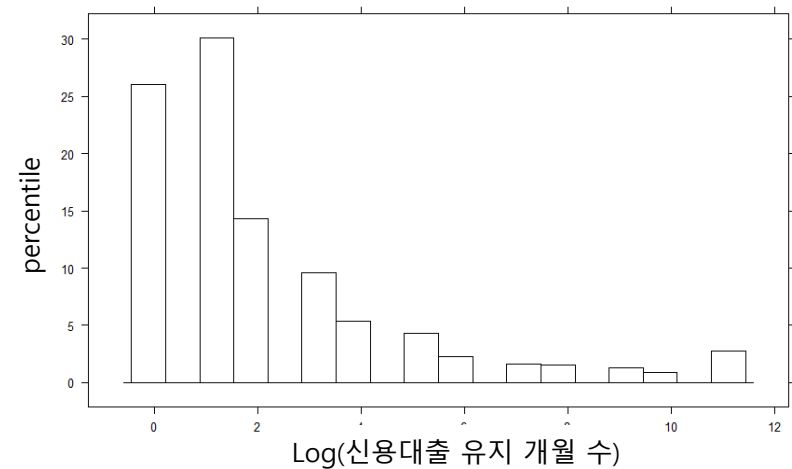
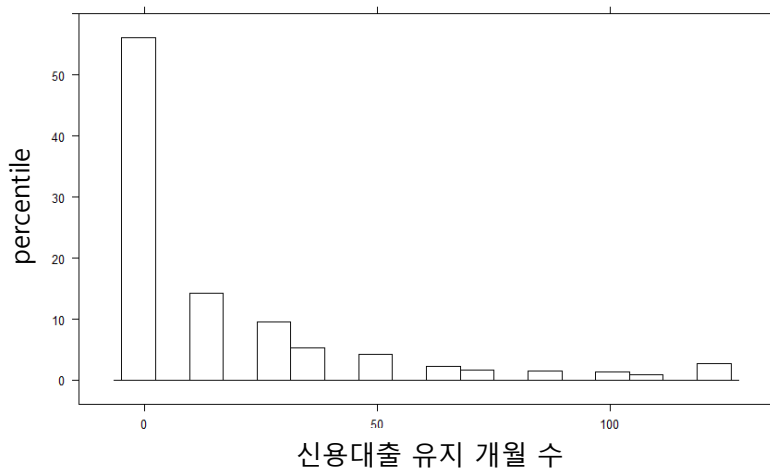
- 금액 변수들의 단위를 '1원'으로 변환

	구분	변수 설명	단위
DATA	SCI	총 대출금액	1,000원
		총 신용대출 금액	
		총 은행대출 금액	
		총 카드사 등 대출 금액	
	한화생명	추정소득	10,000원
		가구추정소득	
		납입보험료	
		.	
		.	
		.	
	SKT	최대월납입보험료	1원
		가입자대출_원	
		납부요금_원	
		단말기가격_원	
		당월연체금액_원	
		년간최대연체금액_원	
		남은할부금_원	

◆ 변수 변환

- 0으로 치우친 변수를 로그 변환을 통해 분산시킴
ex) 총 대출 금액, 가계합산 보장성, 단말기가격...
- 12개월 단위의 유지 개월 수 변수, 순서형 자료로 변환
ex) 신용대출 계좌 유지 개월 수, 2산업분류 계좌 유지 개월 수...

0, 1, 13, 25 $\xrightarrow{\text{Round up}}$ 0, 1, 2, 3



◆ 변수 변환

- 다수의 Factor형 변수 존재
 - 금액이나 건수 등의 변수는 범주로 되어있으나 순서형이므로 범주처리 하지 않음
 - 이를 제외한 아래에 있는 변수들을 범주화

	구분	변수 설명	타입
DATA	한화생명	직업	Factor
		배우자 직업	
	SKT	나이	
		성별	
		멤버쉽등급	
		결합상품가입여부	
		납부방법	
		회선상태	

◆ 변수 변환

- 날짜 데이터와 분기 데이터의 단위 통일
 - 최초 대출 날짜는 월 단위, 가입 년 월_분기는 분기 단위
 - 분기별 특성이 월 단위보다 뚜렷하므로 두 단위를 일치시킴
 - 2016년 2분기 = 0을 기준으로 1991년 3분기까지 각 분기마다 0.25의 값을 줌
ex) 1991.09 -> 1991년 3분기 = 16.75, 1991.12 -> 1991년 4분기 = 16.5

	구분	변수 설명	변환 예시
DATA	한화생명	최초 대출 날짜	ex) 2016년 2분기 → 16.25
	SKT	가입 년 월_분기	

◆ 비식별 / 결측값

• 비식별

- 4개의 변수에서 존재하는 비식별 처리 관측치는 비식별의 특수성을 고려하여 하나의 범주로 처리

	구분	변수 설명	개수
DATA	한화생명	직업	1,189
		배우자 지역	1,027
		나이	430
		성별	430

• 결측값

- 결측값이 존재하는 5개 변수 중, 범주가 적은 변수의 결측값을 MICE 처리 ex) 멤버쉽등급, 납부방법

- 직업, 배우자 직업, 막내 자녀나이의 결측값은 해당 변수의 범주가 많아 하나의 범주로 지정

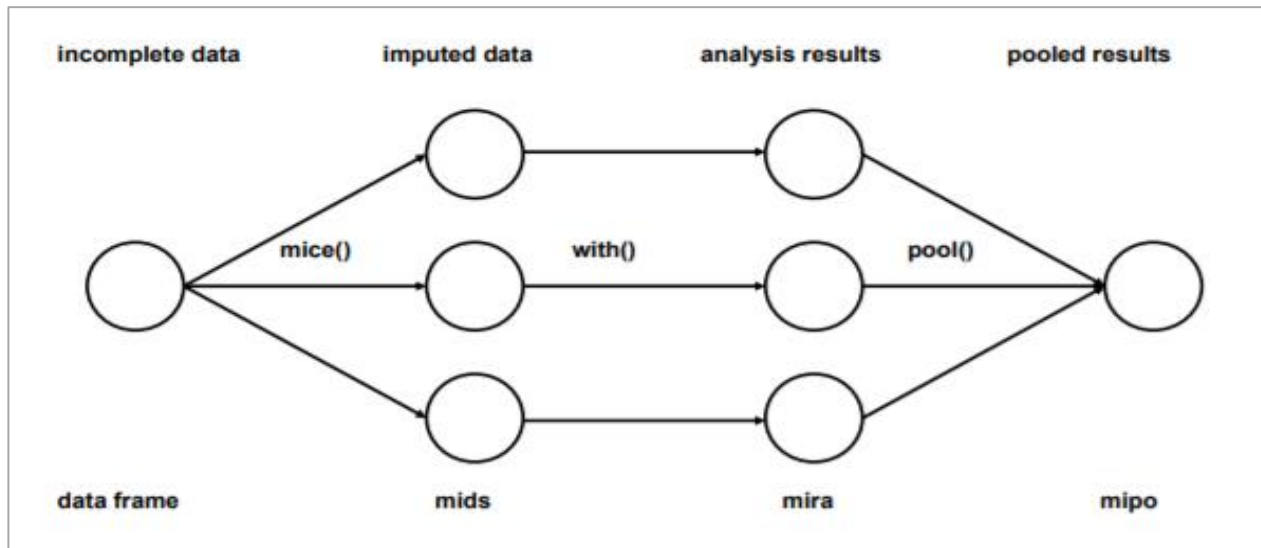
	구분	변수 설명	개수	변환
DATA	한화생명	직업	464	NULL
		배우자 직업	45,709	NULL
		막내 자녀나이	1,027	0
	SKT	멤버쉽등급	46,015	MICE
		납부방법	2,833	MICE

◆ 비식별 / 결측값

- 결측값

- MICE (multiple imputation, chained equations)

- 1) 결측값을 plausible value(mean/ mode...)로 처리
- 2) 결측값이 존재하는 변수를 예측변수로 하는 모형 설정
- 3) 설정된 모형으로 Pooling하여 Imputation 실시



대출상환 예측 알고리즘 개발

- Preprocessing
- Feature value
- Imbalanced data
- Model
- Conclusion

◆ 파생 변수

- Vec.factor

- 대출 기관 대출 유무 (Factor)

은행	캐피탈	2산업	기타
2	1	0	0
3	0	1	1
1	4	2	0
0	3	0	0



은행	캐피탈	2산업	기타	Vec.factor
1	1	0	0	1100
1	0	1	1	1011
1	1	1	0	1110
0	1	0	0	100

- Vec.prop

- 기관별 대출 유무에 대한 비율

Vec.factor
1100
1011
1110
100



Vec.prop
0.5
0.75
0.75
0.25

◆ 파생 변수

- TOT_LNIF_CNT

- 총 대출 건수
- 은행 건수 + 카드/보험/캐피탈 건수
+ 2산업 건수 + 기타 건수

은행	캐피탈	2산업	기타	TOT_LNIF_CN
2	1	0	0	3
3	0	1	1	5
1	4	2	0	7
0	3	0	0	3

- PropBNK.SPART

- 은행과 2산업의 대출 건수 중 은행 비율
- 은행건수/(은행건수+2산업건수)

은행	캐피탈	2산업	기타	PropBNK.SPART
1	1	0	0	1
1	0	1	1	0.5
1	1	1	0	0.5
0	1	0	0	0

◆ 파생 변수

- Diff_occrr_mdif

- 신용대출과 2산업대출 건수 차이
- (신용대출 건수 - 2산업대출 건수)

은행	캐피탈	2산업	기타	Diff occrr Mdif
2	1	0	0	2
3	0	1	1	2
1	4	2	0	-1
0	3	0	0	0

- Is.same.crdt_sptct

- 신용대출자와 신용대출 미 이용자 구분
- Diff_occrr_mdif = 0
(신용 대출 건수 = 2산업 대출 건수)
신용 대출 미 이용자와 은행과 2산업의 대출
건수가 동일한 사람을 구별하기 위함

은행	캐피탈	2산업	기타	Is.same. crdt sptct
1	1	0	0	1
1	0	1	1	1
1	1	1	0	1
0	1	0	0	0

신용대출자 = 0, 신용대출 미 이용자 = 1

◆ 파생 변수

- diff.card.grad

- 최초 신용등급과 최근 신용등급의 차이
- (최초 신용등급- 최근 신용등급)

최초 신용등급	최근신용등급	diff.card.grad
2	1	1
2	6	-4
1	1	0
0	0	0

- is.strt.cg.0

- 최초 신용등급이 0인 경우와 그렇지 않은 경우 구분
- 최초 신용등급이 존재하지 않았던 사람이 최근 신용등급이 생겨서 발생한 값과 최초 신용등급이 존재하지만 최근 신용등급의 차이가 발생한 사람과의 구분을 위한 변수

최초 신용등급	최근신용등급	diff.card.grad	is.strt.cg.0
2	1	1	0
2	6	-4	0
0	3	-3	1
0	0	0	1

◆ 파생 변수

- Diff_repy.crln

- 한화생명신용대출금액 반환 값
- (한화생명신용대출금액 - 한화생명신용상환금액)

한화생명신용대출금액	한화생명신용상환금액	Diff_repy.crln
10000000	3000000	7000000
13000000	10000000	3000000
0	0	0
70000000	70000000	0

- Is.same.repy.crln

- 한화생명신용대출이용자와 미 이용자 구별 변수
- 한화생명신용대출을 전부 상환한 고객과 한화생명에서 신용대출을 받은 기록이 없는 고객을 구별하기 위함 (신용대출이용자 = 1, 미 이용자 =0)

한화생명신용대출금액	한화생명신용상환금액	Diff_repy.crln	Is.same.repy.crln
10000000	3000000	7000000	1
13000000	10000000	3000000	1
0	0	0	0
70000000	70000000	0	1

◆ 파생 변수

- Diff_repy.crln

- 한화생명신용대출금액 반환 값
- (한화생명신용대출금액 - 한화생명신용상환금액)

한화생명신용대출금액	한화생명신용상환금액	Diff_repy.crln
10000000	3000000	7000000
13000000	10000000	3000000
0	0	0
70000000	70000000	0

- Is.same.repy.crln

- 한화생명신용대출이용자와 미 이용자 구별 변수
- 한화생명신용대출을 전부 상환한 고객과 한화생명에서 신용대출을 받은 기록이 없는 고객을 구별하기 위함 (신용대출이용자 = 1, 미 이용자 =0)

한화생명신용대출금액	한화생명신용상환금액	Diff_repy.crln	Is.same.repy.crln
10000000	3000000	7000000	1
13000000	10000000	3000000	1
0	0	0	0
70000000	70000000	0	1

대출상환 예측 알고리즘 개발

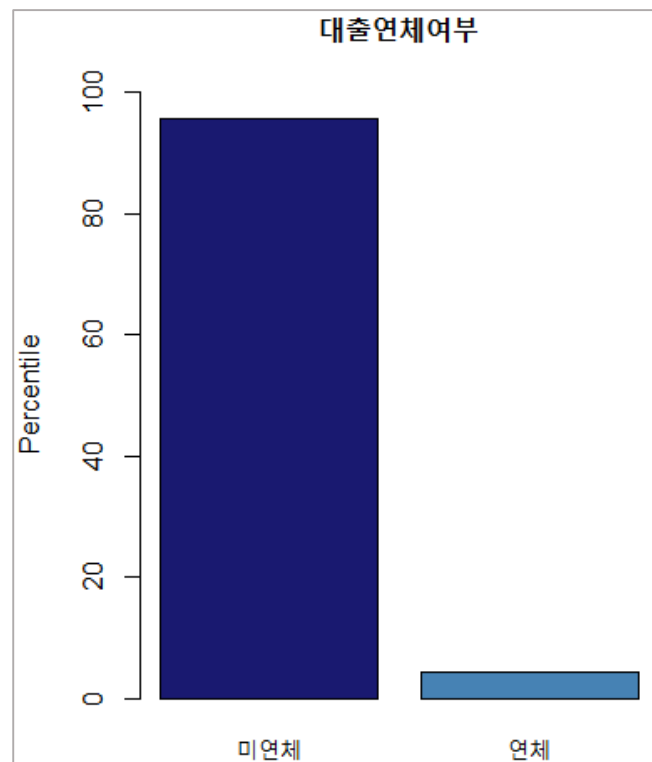
- Preprocessing
- Feature value
- Imbalanced data
- Model
- Conclusion

◆ 대출연체여부 빈도 확인

- Imbalanced Data

- 총 100,234명의 대출연체여부는 미연체 95,946명, 연체 4,287명으로 구성되어 있음
- 대출연체여부가 미연체에 크게 편향되어 있음
- 빈도가 높은 케이스에 과적합하여 빈도가 낮은 케이스에 대한 예측 정확도가 낮아짐

→ Data Cleansing을 통해 연체자 데이터 복제

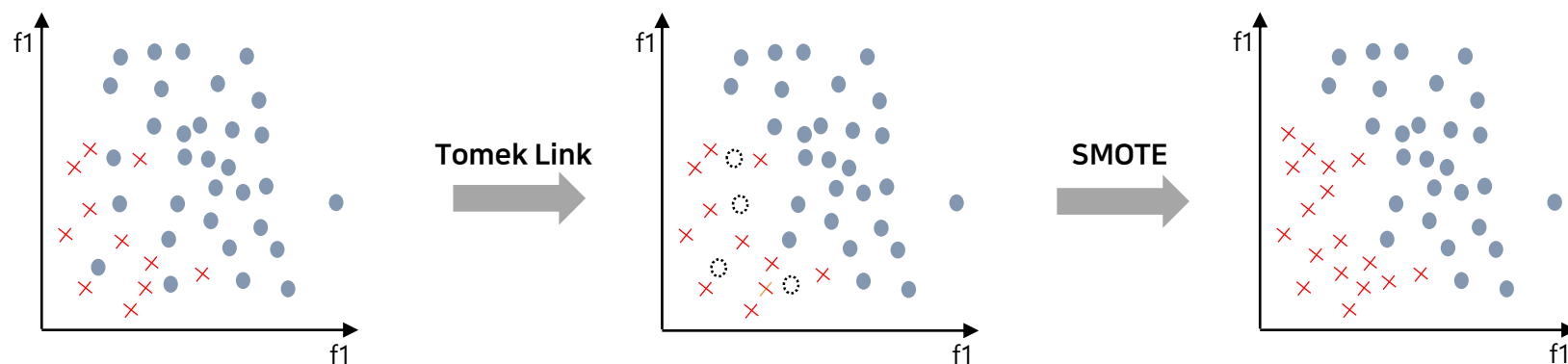


◆ Data Cleansing

• 데이터 복제

- 데이터 합성을 통해 소수 케이스인 연체자를 새로 합성
- 연체자 중, 미연체자와 분류가 어려운 데이터로 합성할 경우, 학습 성능이 저하됨

→ Tomek Link로 overlapping을 제거한 뒤, SMOTE로 연체 데이터 복제



◆ Data Cleansing

- Tomek Link

- Tomek Link 정의

$x \in S_{\min}(\text{minority of } S), y \in S_{\max}(\text{majority of } S)$

$d(x, z) < d(x, y) \text{ or } d(x, y) > d(y, z)$

(x, y) is Tomek Link

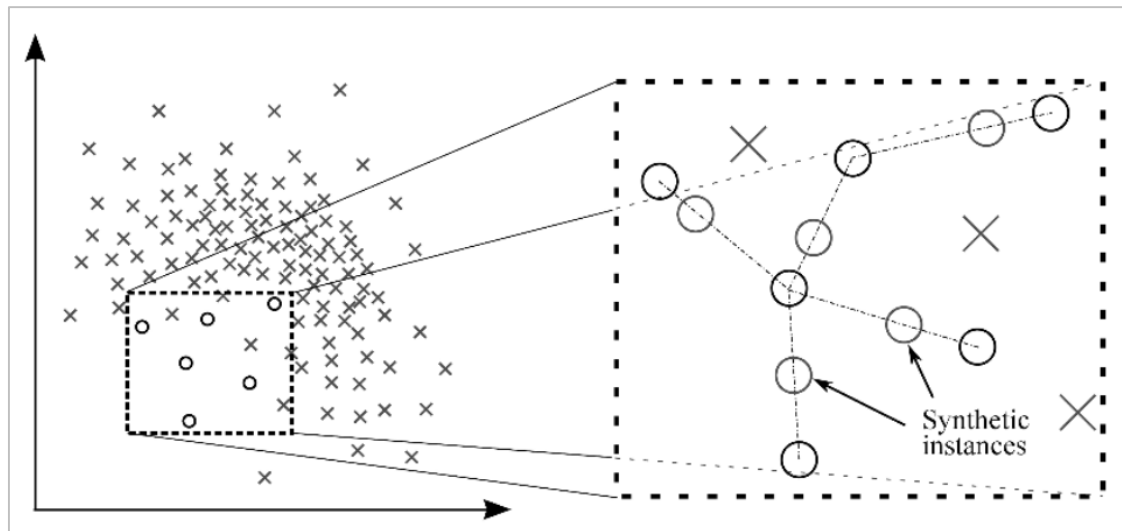
- Tomek Link를 형성한다면 두 샘플 중, 하나는 noise이거나 둘 다 분류의 경계에 있는 것으로 볼 수 있으므로, 모든 쌍이 같은 범주에 있을 때까지 Tomek Link를 제거한다.
 - Overlapping을 제거하여 잘 정의된 데이터를 얻을 수 있고, 이는 개선된 성능으로 이어진다.

◆ Data Cleansing

- Synthetic Minority Oversampling Technique

- 소수 범주의 데이터를 서로 보간하여 새로운 인공적인 데이터 합성

- 1) Minority 범주 데이터를 sample로 취하여, k-nearest neighbor를 찾음
- 2) sample과 이웃 간의 차를 구함
- 3) 이 차에 0~1사이의 값을 곱하여 원래 sample에 추가함



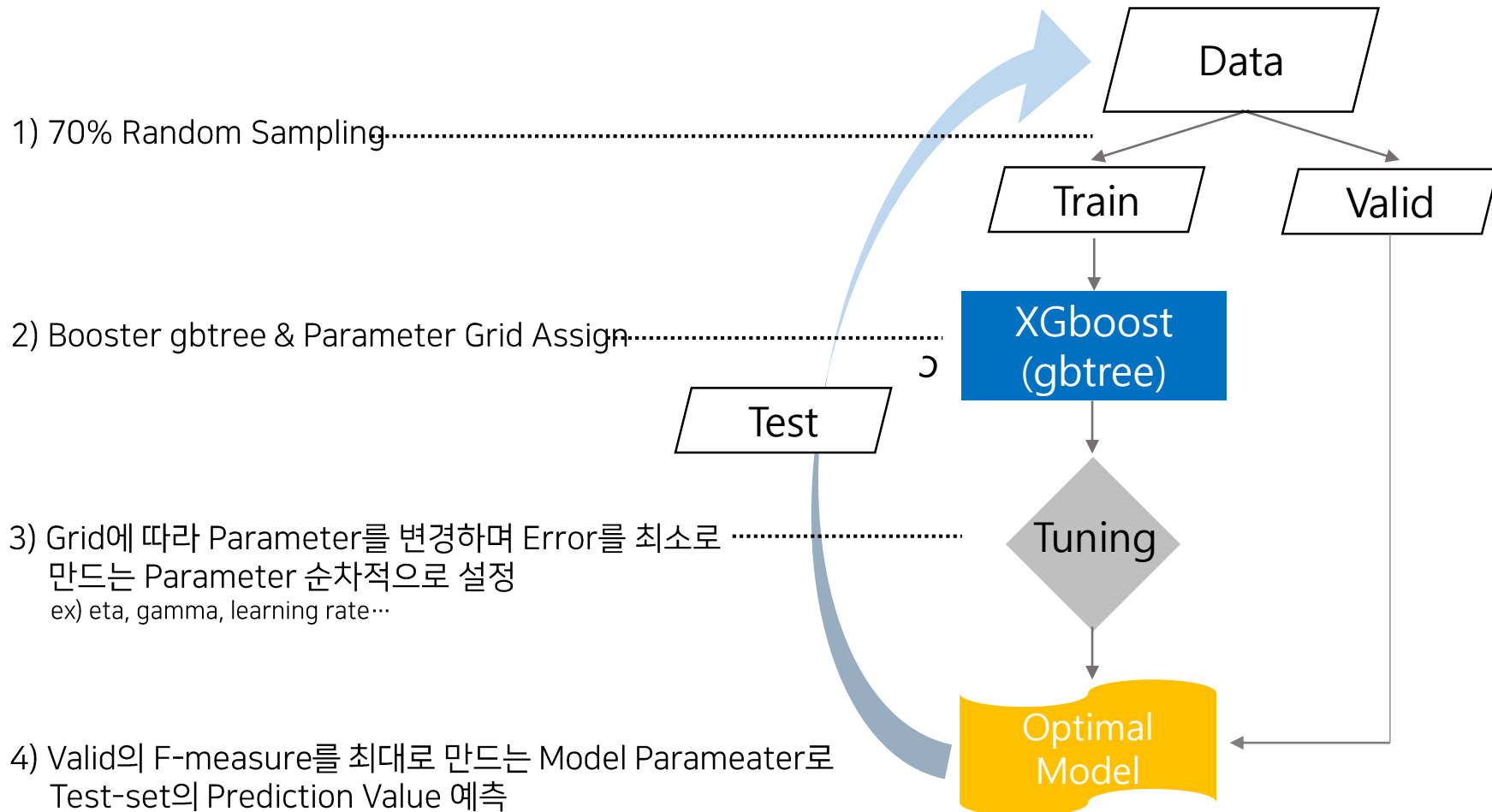
대출상환 예측 알고리즘 개발

- Preprocessing
- Feature value
- Imbalanced data
- Model
- Conclusion

◆ XGboost

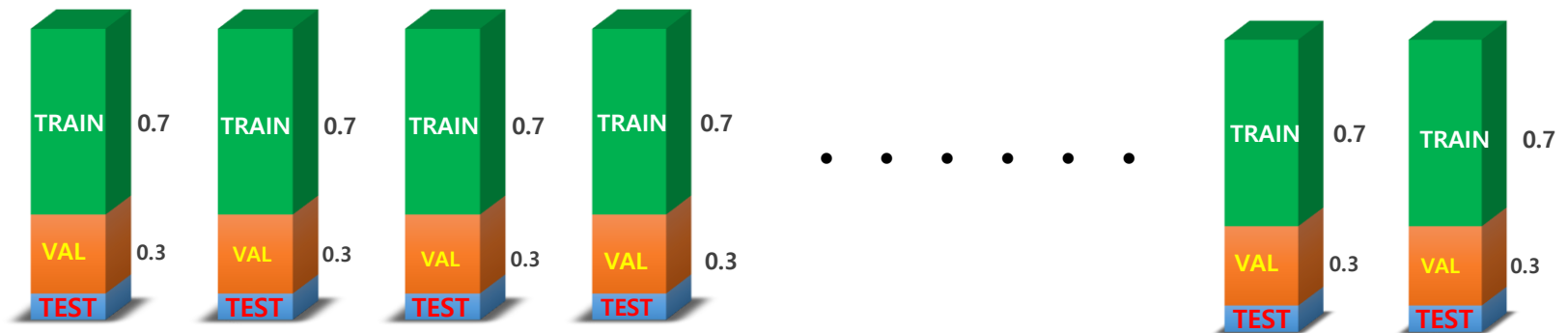
- Gradient boosting 알고리즘 사용
 - ✓ 이는 이전 모델에서의 잔차를 줄이기 위해서 Parameter를 Grid Search를 통해 순차적으로 개선함으로써 최적의 Parameter를 찾아 모델의 정확도를 높임
- 빠른 모델 수행
 - ✓ Tree를 구성할 때 노드 단위로 병렬처리를 하여 계산 속도가 빠르므로 Bagging algorithm 적용 용이
- 높은 performance
 - ✓ xgboost는 각 변수에 알맞게 정규화(Regularized)하므로 더 높은 정확도 산출과 과적합을 방지하고자함

◆ XGboost



◆ Bagging algorithm


- Xgboost Model Fitting
 - ✓ Random Sampling한 Train set으로 Valid set의 F-measure를 최적으로 만드는 Xgboost 모델 생성을 350회 반복



◆ Bagging algorithm

- Prediction Voting
 - ✓ 350개의 Xgboost 모델에 Test-set을 적합하여 Prediction의 빈도를 비율로 변환하고 Voting함으로써 Overfitting을 방지 및 분산 최소화를 통해 예측력을 높임

Model ID \	Xgboost1	Xgboost2	Xgboost3	Xgboost350	Target
100	0	0	0		0	0
188	0	0	1		0	0
⋮						
102204	1	0	1		1	1



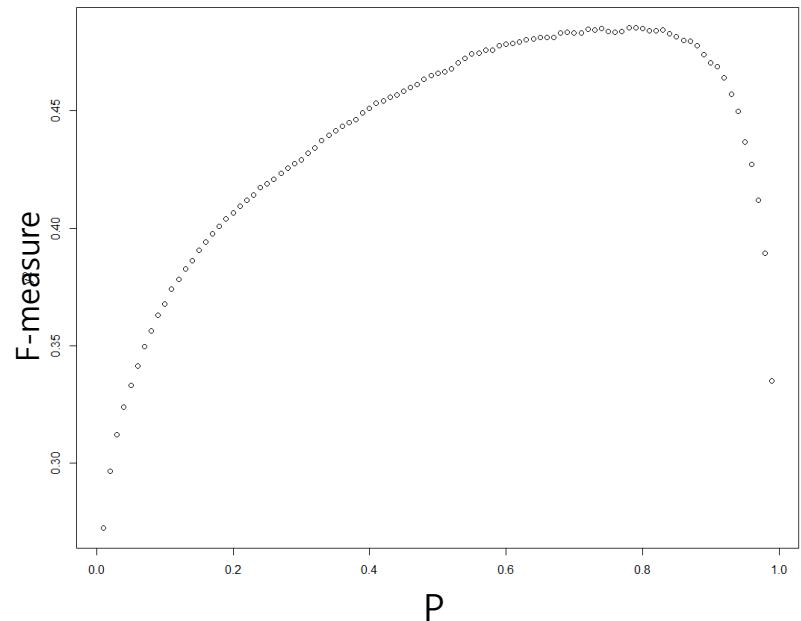
대출상환 예측 알고리즘 개발

- Preprocessing
- Feature value
- Imbalanced data
- Model
- Conclusion

◆ 모델 성능

- 신용평가척도 모델 다변화
 - 한화생명, SKT의 데이터를 결합한 Alternative 신용평가모델을 통하여 신용평가 척도 모델을 다변화
- 잠재적 고객 평가
 - 이를 통해 기존의 신용평가모델로는 대출 여부의 판단이 부정확한 중금리 시장의 잠재적인 고객들 평가를 통해 중금리 시장 확대 및 고객 창출
- 채무 부담 경감
 - 사회적으로는 중금리 대출을 받지 못해, 2산업, 사채 등에서 고금리로 대출을 받는 중신용자들의 채무 부담 경감

평균	48.5%
최적의 p값	0.78
350번 시뮬레이션을 돌렸을 때 평균	48%
95%신뢰구간	[47.4, 48.6]



THANK YOU FOR YOUR ATTENTION