
antock

Agenda

기술 검증 추진 배경

공공 데이터 유의성 검증

신용평가 (예측) 모형 검증

현장 테스트 결과

기술 검증 개요

당사는 기업 빅 데이터 AI 분석 기술을 주제로 IBK 중소기업은행 1st Lab 프로그램에 선정되었으며,
리스크 총괄 부서의 요청으로 대안 신용평가 적용 가능성에 대한 기술 검증을 수행하게 되었습니다.

IBK 중소기업은행 1st Lab 논의

IBK 중소기업은행 1st Lab PoC 논의

(22. 01. 05 (수) 16:00 - 현재)

참! 좋은 은행 IBK기업은행

- **논의 배경**
  - 당사는 IBK 중소기업은행 1st Lab 프로그램에 선정되어 보유 기술력에 대한 기술 검증을 진행하기로 합의
- **논의 내용**
  - 당사가 제시한 3개 기술검증 주제 (ESG 평가, 혁신 기술기업 발굴, 대안 신용평가) 중 리스크 총괄부 후자 택일
  - 이에 따라, 구체적인 기술 검증 방안에 대한 구체적 설계 및 이에 대한 당사 관점의 데이터 준비 및 실험 진행

세부 논의 내용

대안 신용 평가 관점에서의 당사 기술력 유효성 검증

검증 목적	• 기존 신용평가 모형의 경우, 재무 데이터가 부재하거나 유의미하지 않은 중소기업에 대한 예측 변별력 문제 존재 • 이에 따라, 대안 데이터 (Alternative Data) 관점에서, 비재무 정보를 근거로 유의미한 분석과 예측이 가능한지 실험
검증 범위	• 당사가 보유한 비재무 대안 데이터 라이브러리 내 주요 항목 별 기업 부실 (리스크 수준) 간의 연관관계 분석 • 또한, 데이터 조합 및 인공지능 분석을 통해 실질적으로 중소기업에 대한 리스크 예측 적용 실효성을 탐색
검증 방안	• IBK 중소기업은행 실질 대출 차주 기업 및 부도 여부에 대한 학습 데이터를 이관하여 데이터 유의성 및 일차적인 모형 개발 수행 • 또한, IBK 기업은행이 보유한 검증 데이터에 대한 개발 모형을 적용하여 이에 대한 내부적인 검수 진행

기술 개발 가설

당사는 본 기술 검증과 관련하여, 기업 별 공공 데이터 발생 양상을 분석함으로써, 개별 기업의 경영 성과를 예측할 수 있다는 가설을 수립하게 되었습니다.

주요 가설 및 기술 개발 배경

"기업 별 공공 데이터의 발생 양상과 경영 리스크에는 유의미한 연관관계가 존재할 것이다."

1 공공 데이터

- 공공 데이터는 특정 산업 및 유형을 막론하고 국내 중소기업에 대한 정보를 포괄적으로 제공
- 또한, 정부 / 공공을 원천으로 하는 신뢰도 높은 데이터로서 안정적인 수급이 가능

2 데이터 발생

- 실제 데이터의 세부 내용 외, 데이터 발생 로그를 파생 데이터로 가공하여 활용
- 공공 데이터의 발생 여부, 발생 시점, 발생 주기 등의 원초적 정보가 기업 현황의 직간접적인 단서로 작용

3 경영 리스크

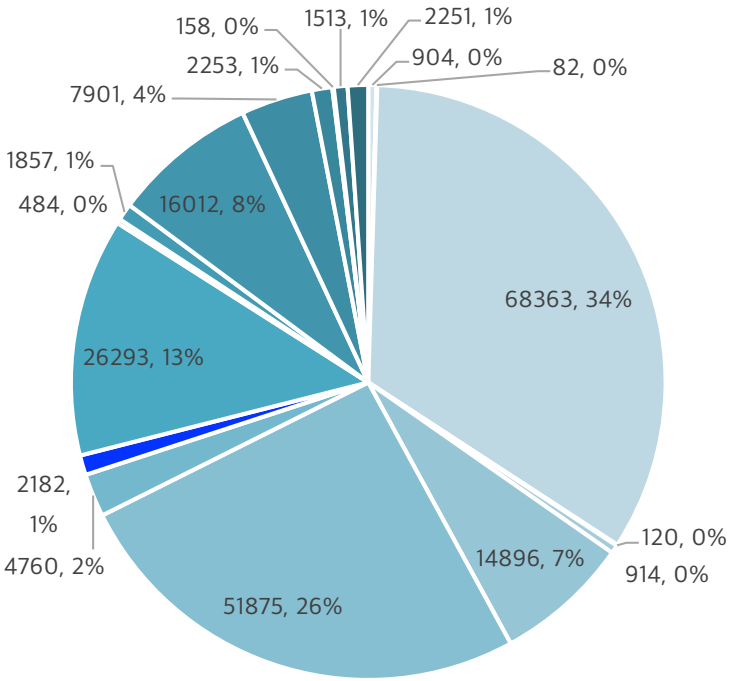
- 개별 기업 관점에서의 주요 실적 악화의 징후를 포착하며, 이에 따른 경쟁력 변화 평가
- 금융 기관 차원에서는 여신 연체, 부도 등을 의미하며, 투자자의 경우 가치 하락, 회수 위험 등

IBK 중소기업은행 1st Lab 프로그램을 통해 당사 수립 가설에 대한 일차적인 검증 수행

1 가설 수립 배경 : 공공 데이터 의의

공공 데이터를 활용하여 보다 포괄적인 기업군을 진단할 수 있고, 이는 객관적인 평가 / 분석의 토대가 되며, 안정적인 정보 수급이 가능합니다.

IBK 중소기업은행 학습 표본 업종 분포



업종	비중
A	0.4%
B	0.0%
C	33.7%
D	0.1%
E	0.5%
F	7.3%
G	25.6%
H	2.3%
I	1.1%
J	13.0%
K	0.2%
L	0.9%
M	7.9%
N	3.9%
P	1.1%
Q	0.1%
R	0.7%
S	1.1%

데이터 관련 주요 시사점

- **민간 데이터 업종 편향성**
  - ➢ 민간 데이터 연계 시, 제공 주체에 따른 업종 제한 적용이 필연적이며, 이는 분석 범위를 현격히 축소
    예시) '배달의 민족' 데이터는 전체 1% 수준의 기업 표본에만 적용 가능
- **민간 데이터 정보 신뢰성**
  - ➢ 민간 데이터 생성 및 가공에 대한 자체적인 기준이 존재함에 따라, 이에 대한 파악이 용이하지 않은 구조
    설명) 모형에 대한 객관적인 검증 시, 활용 데이터에 대한 설명 한계 발생
- **민간 데이터 의존 위험성**
  - ➢ 또한, 민간 데이터의 경우 수익성 등의 경제 논리 혹은 경쟁 발생 시, 데이터에 대한 공급 중단 가능
    설명) 한국평가데이터 (한국기업데이터) 경우, 자체 제공 서비스 경쟁 시 데이터 차단

2 가설 수립 배경 : 데이터 발생 개념

공공 데이터 발생 로그는 데이터의 세부 내용이 뿐만이 아닌 데이터 항목 자체의 발생 여부, 시점, 그리고 총량 등을 분석하는 개념입니다.

공공 데이터 RFM 모델의 정의

공공 데이터를 최신성(R, Recency), 빈도 (F, Frequency), 규모 (M, Magnitude) 관점에서 분석하는 기법

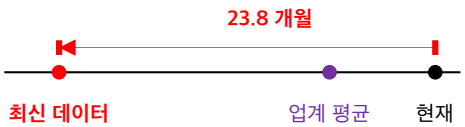
Recency (최신성)

- 중소기업 공공데이터가 마지막으로 발생한 시점과 현재와 기간 차 산출

➢ **현재 관점에서 기업 별 사업 활동의 유효성 검증**

Case 1 : 기업 공공 데이터, **최근에 발생한 경우**

Case 2 : 기업 공공 데이터, **장기간 부재한 경우**

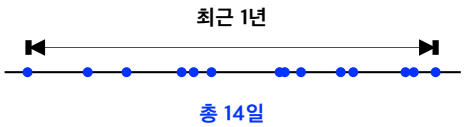


Frequency (빈도)

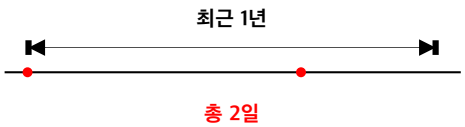
- 특정 기간 동안 개별 기업 공공 데이터의 발생 개수 계산

➢ **기업의 사업의 활동의 지속/영속성 등을 정량적으로 도출**

Case 1 : 기업 공공 데이터, **자주 발생 된 경우**



Case 2 : 기업 공공 데이터, **드물게 발생 된 경우**



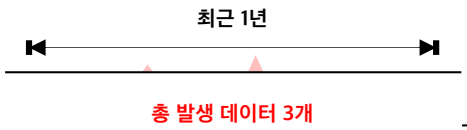
Magnitude (규모)

- 특정 기간 동안 기업 별 공공 데이터 발생 총량 (합산) 산출

➢ **시장 내에서 개별 기업의 사업 활동 영향력 / 파급력 확인**

Case 1 : 기업 공공 데이터, **발생량이 많은 경우**

Case 2 : 기업 공공 데이터, **발생량이 적은 경우**



사업성 양호 신호

리스크 관리 신호

2 가설 수립 배경 : 기업 리스크 연계

실제 귀납적으로 데이터 발생에 대한 정보를 실질적인 기업 사업 수행 현황과 연계한 결과, 유의미한 상관 관계가 관찰되었습니다.

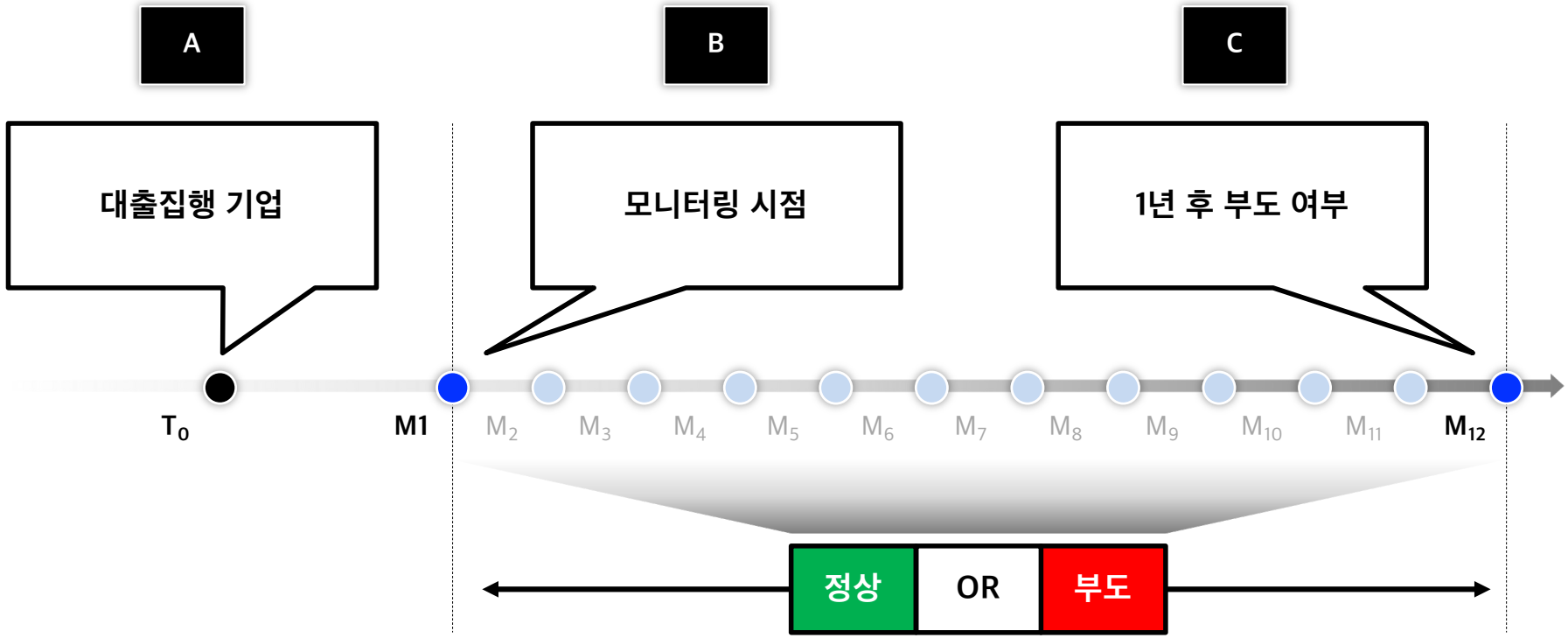
공공 데이터 예시 1	공공 데이터 예시 2	공공 데이터 예시 3	공공 데이터 예시 4	…
• 국민연금공단	• 특허청	• 벤처기업협회	• 고용노동부	
NPS 국민연금	KIPRIS	벤처기업확인기관 벤처확인종합관리시스템	WORKNET	
고용 정보가 1년간 누락된 경우 :	최근 특허 출원이 350% 증가 :	벤처기업 인증 갱신 정보 미 발생 :	채용 공고 데이터 3일전 발생 :	
"확인"	"확인"	"확인"	"확인"	
"최근 코로나 여파로 매출이 큰 폭으로 감소하여, 직원들이 되거 퇴사를 …"	"신규 투자 유치에 성공하여, 기술 개발이 원활하게 진행되어 특허 출원을 …"	"지금 인증에 대한 갱신을 진행할 수 있는 사업 측면의 여유가 전혀 없는 …"	"큰 프로젝트를 수주하게 되어서 사업 확장을 위한 채용을 활발하게 …"	

기술 검증 설계

당사는 본 가설을 검증하기 위한 작업으로, IBK 중소기업은행 측이 제고한 차주 표본에 대한 모니터링 시점 별 1년 후 부도 여부에 대한 공공 데이터 기반의 예측 시뮬레이션을 진행하게 되었습니다.

IBK 중소기업은행 1st Lab 성능 평가 관점

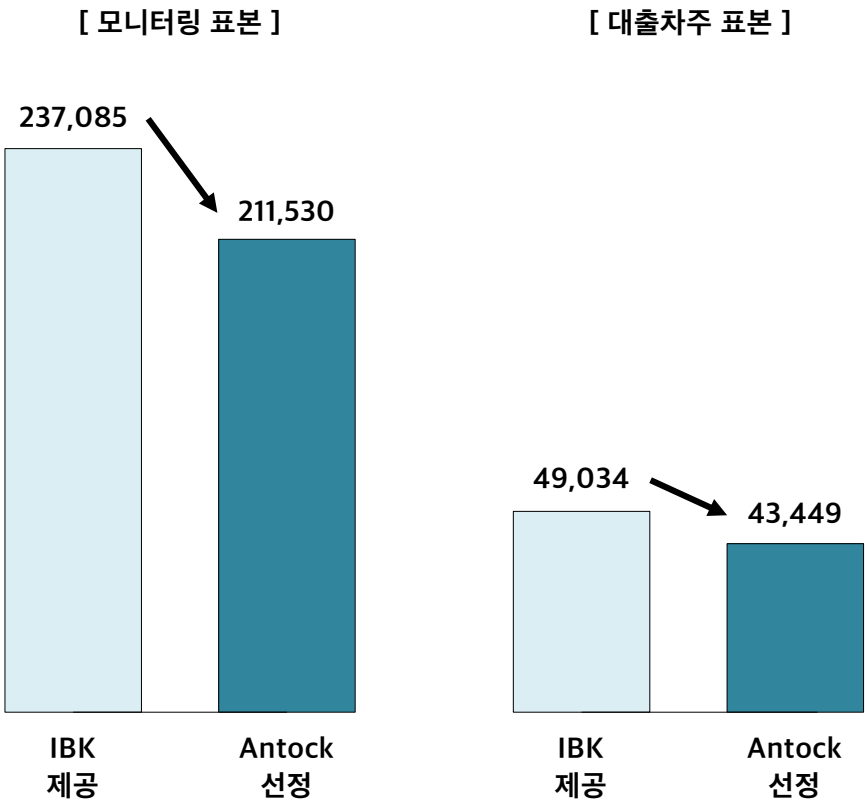
IBK 대출 차주에 대한 모니터링 시점 기준으로 1년 후 부도 여부를 예측할 수 있는가?



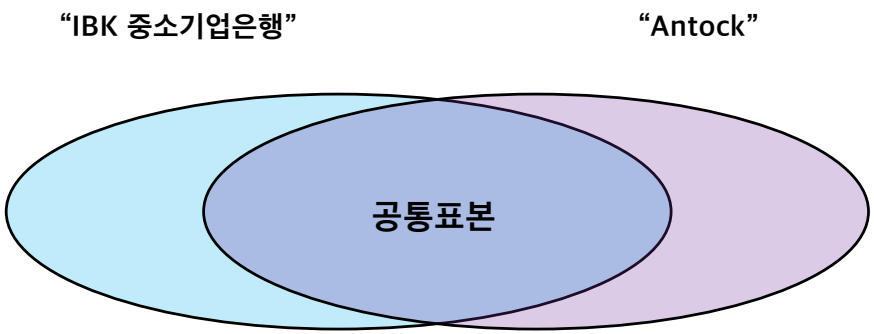
A 기술 검증 설계 : 대상 표본

IBK 중소기업은행 측으로부터 총 237,085개 (49,034개 기업) 표본을 제공 받았으며, 당사 시스템 상에서 식별되는 211,530개 (43,449개 기업) 표본을 시뮬레이션 대상으로 최종 결정하였습니다.

IBK 1st Lab 제공 표본 및 선정 결과



시뮬레이션 대상 표본 감소 원인

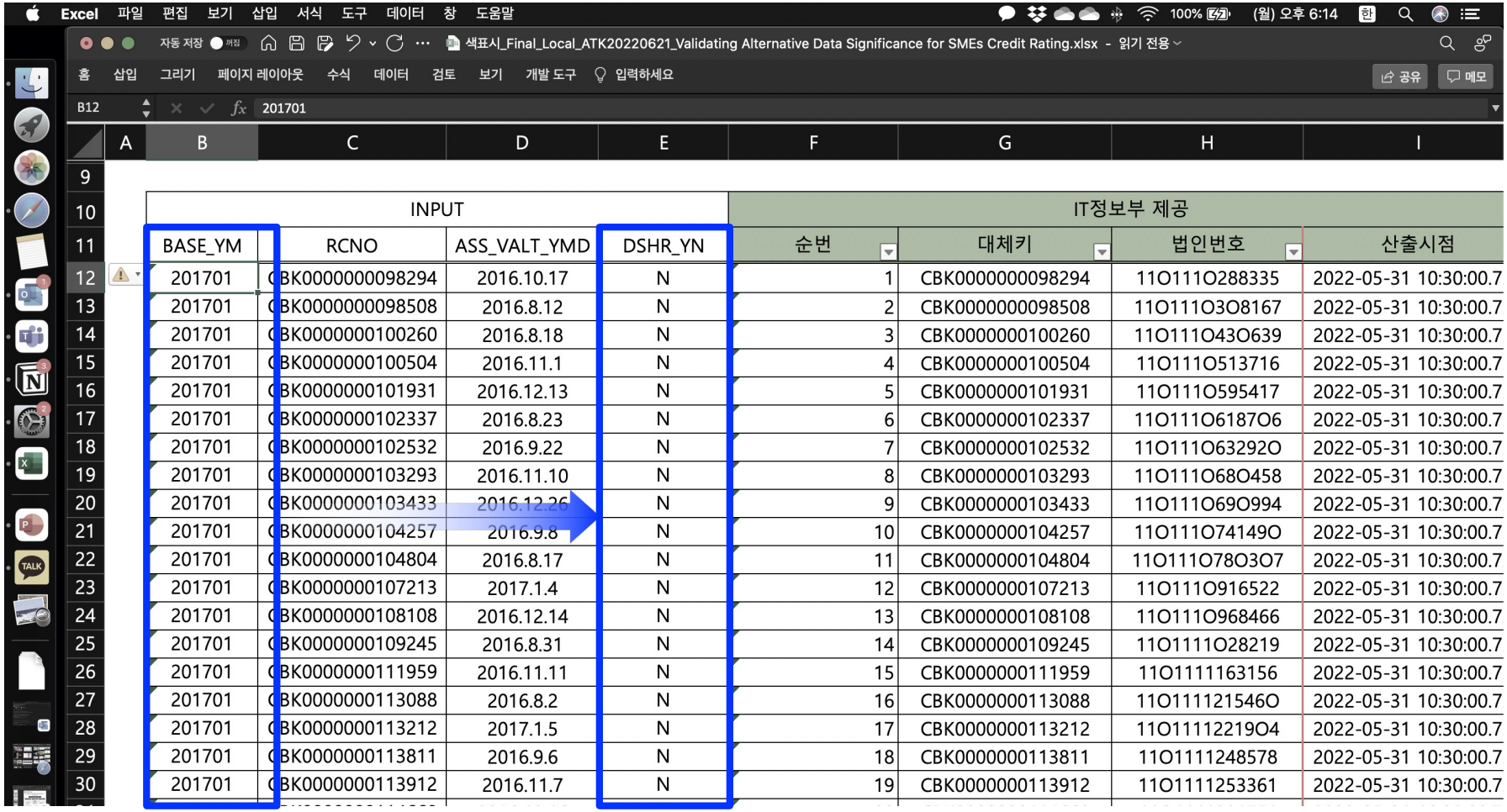


- IBK 중소기업은행 대출 차주 기업 표본 중 당사 시스템 내 공통적으로 등록된 표본만이 시뮬레이션 대상으로 적합 (데이터 수집 차원)

➢ **현재 당사 Hubble Database 2.0 Alpha 안정화 이후에는 식별 누락 기업 표본의 수가 현격하게 감소할 것으로 추정**

B 기술 검증 설계 : 예측 시점

당사는 대출이 실행된 기업 차주의 모니터링 시점 (BASE_YM) 별 1년 후 부도 여부를 기준으로 시뮬레이션을 진행하기로 합의하였습니다.



INPUT				IT정보부 제공			
BASE_YM	RCNO	ASS_VALT_YMD	DSHR_YN	순번	대체키	법인번호	산출시점
201701	CBK0000000098294	2016.10.17	N	1	CBK0000000098294	11O111O288335	2022-05-31 10:30:00.7
201701	CBK0000000098508	2016.8.12	N	2	CBK0000000098508	11O111O3O8167	2022-05-31 10:30:00.7
201701	CBK0000000100260	2016.8.18	N	3	CBK0000000100260	11O111O43O639	2022-05-31 10:30:00.7
201701	CBK0000000100504	2016.11.1	N	4	CBK0000000100504	11O111O513716	2022-05-31 10:30:00.7
201701	CBK0000000101931	2016.12.13	N	5	CBK0000000101931	11O111O595417	2022-05-31 10:30:00.7
201701	CBK0000000102337	2016.8.23	N	6	CBK0000000102337	11O111O6187O6	2022-05-31 10:30:00.7
201701	CBK0000000102532	2016.9.22	N	7	CBK0000000102532	11O111O632920	2022-05-31 10:30:00.7
201701	CBK0000000103293	2016.11.10	N	8	CBK0000000103293	11O111O68O458	2022-05-31 10:30:00.7
201701	CBK0000000103433	2016.12.26	N	9	CBK0000000103433	11O111O69O994	2022-05-31 10:30:00.7
201701	CBK0000000104257	2016.9.8	N	10	CBK0000000104257	11O111O74149O	2022-05-31 10:30:00.7
201701	CBK0000000104804	2016.8.17	N	11	CBK0000000104804	11O111O78O3O7	2022-05-31 10:30:00.7
201701	CBK0000000107213	2017.1.4	N	12	CBK0000000107213	11O111O916522	2022-05-31 10:30:00.7
201701	CBK0000000108108	2016.12.14	N	13	CBK0000000108108	11O111O968466	2022-05-31 10:30:00.7
201701	CBK0000000109245	2016.8.31	N	14	CBK0000000109245	11O1111O28219	2022-05-31 10:30:00.7
201701	CBK0000000111959	2016.11.11	N	15	CBK0000000111959	11O1111163156	2022-05-31 10:30:00.7
201701	CBK0000000113088	2016.8.2	N	16	CBK0000000113088	11O1111215460	2022-05-31 10:30:00.7
201701	CBK0000000113212	2017.1.5	N	17	CBK0000000113212	11O11112219O4	2022-05-31 10:30:00.7
201701	CBK0000000113811	2016.9.6	N	18	CBK0000000113811	11O1111248578	2022-05-31 10:30:00.7
201701	CBK0000000113912	2016.11.7	N	19	CBK0000000113912	11O1111253361	2022-05-31 10:30:00.7

C 기술 검증 설계 : 부도 정의

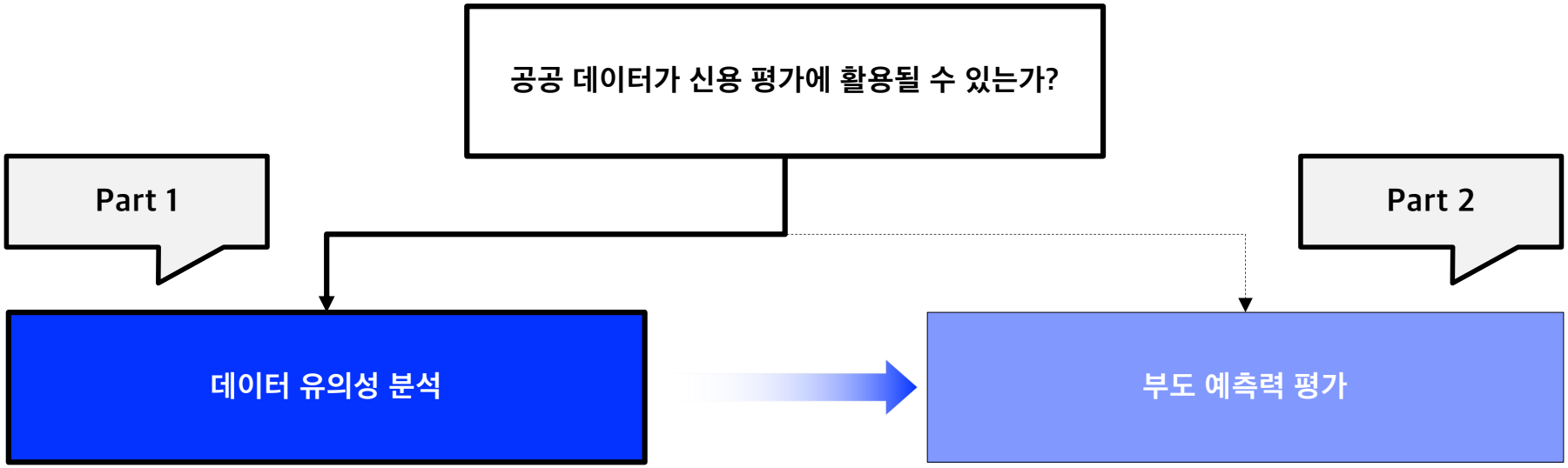
시뮬레이션에서 개별 차주에 대한 부도 여부 판정은 IBK 중소기업은행 측에서 제공한 25개 세부 항목 기준으로 정의되었습니다.

90일 이상 연체 법인	90일 이상 연체 개인	90일 이상 기업 카드	90일 이상 개인 카드	무수익 여신
(01)	(02)	(03)	(04)	(05)
당행어음수표최종부도	특정대손충당적립	일반 상각	카드 상각	채권 매각
(09)	(07)	(08)	(09)	(10)
ABS 발행	회사 정리	화의 부도 사유 여부	기업 워크아웃	개인신용회복지원
(11)	(12)	(13)	(14)	(15)
개인채무회생	파산	법적절차 경매 착수	타기관 부도	기업카드연체카드론대환
(16)	(17)	(18)	(19)	(20)
자본시장본부등록	기타특수채권편입	개인카드연체카드론대환	폐업	개인프리워크아웃
(21)	(22)	(23)	(24)	(25)

기술 검증 방식

보다 구체적으로, 개발 기술에 대한 검증은 기업 부도에 대한 공공데이터 유의성 및 공공데이터 기반한 신용 평가 예측 모델의 실효성 두 가지 형태로 진행되었습니다.

대안 신용평가 PoC 수행 방식



- 당사가 Hubble Database Data Library 내 보유한 다양한 대안 (공공) 데이터를 추출하여 부도 여부 기업 별 통계학적 관점의 차이 / 상관관계 존재하는지 분석

➢ 데이터 공급 업체 측면에서, 신용평가에 활용될 수 있는 주요 공공 데이터 항목을 선별하며 이에 대한 실효성 검증

- 이전 단계에서 검증한 주요 공공 데이터 항목을 결합 / 조합 후 인공지능 모델에 투입하여 실질적인 부도 여부 (신용평가) 예측 모델 개발하여 설명력 검증

➢ 인공지능 분석 업체 측면에서, 실제 현장에서 활용될 수 있는 평가 모델이 구현 가능한지 실험

Agenda

기술 검증 추진 배경

공공 데이터 유의성 검증

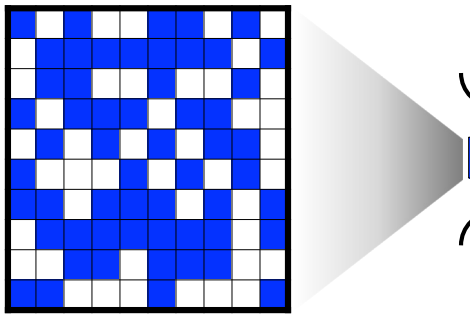
신용평가 (예측) 모형 검증

현장 테스트 결과

데이터 유의성 검증 개요

기업 신용평가에 대한 공공 데이터 유의성 검증을 수행하기 위하여, 당사가 보유한 데이터 라이브러리 내 주요 데이터를 선별하고 시계열 관점의 전처리를 통한 집단 비교 분석을 실시하였습니다.

가 공공 데이터 선별

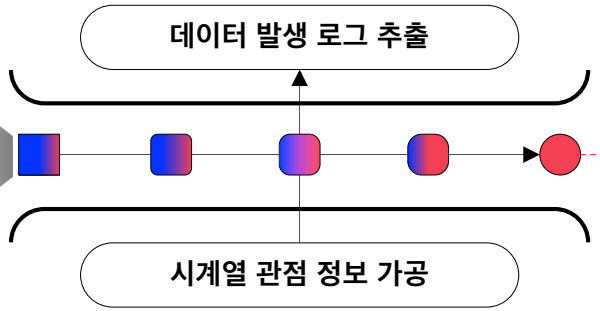


- 당사가 보유한 300가지 이상의 대안 데이터 중, 본 기술 검증에 활용할 핵심 항목을 주요 가이드라인에 따라 선별

➢ 리스크 연관성, 구조적 부합성, 전처리 용이성 등을 고려하여 최종 항목 결정

약 100대 항목 우선적으로 선별

나 전처리 가공 수행

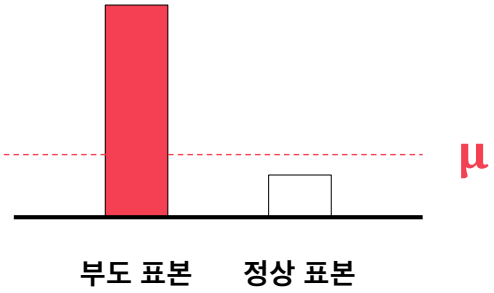


- 선별된 주요 공공 데이터 항목 별로 발생 로그를 추출하고, 심사 기준 시점에 대한 시계열 관점의 변환 수행

➢ IBK 1st Lab 제시한 평가 기준 (모니터링 시점) 일치하는 형태 데이터 전환

2주간 가공 작업 기반 학습 데이터 마련

다 부도 유의성 분석



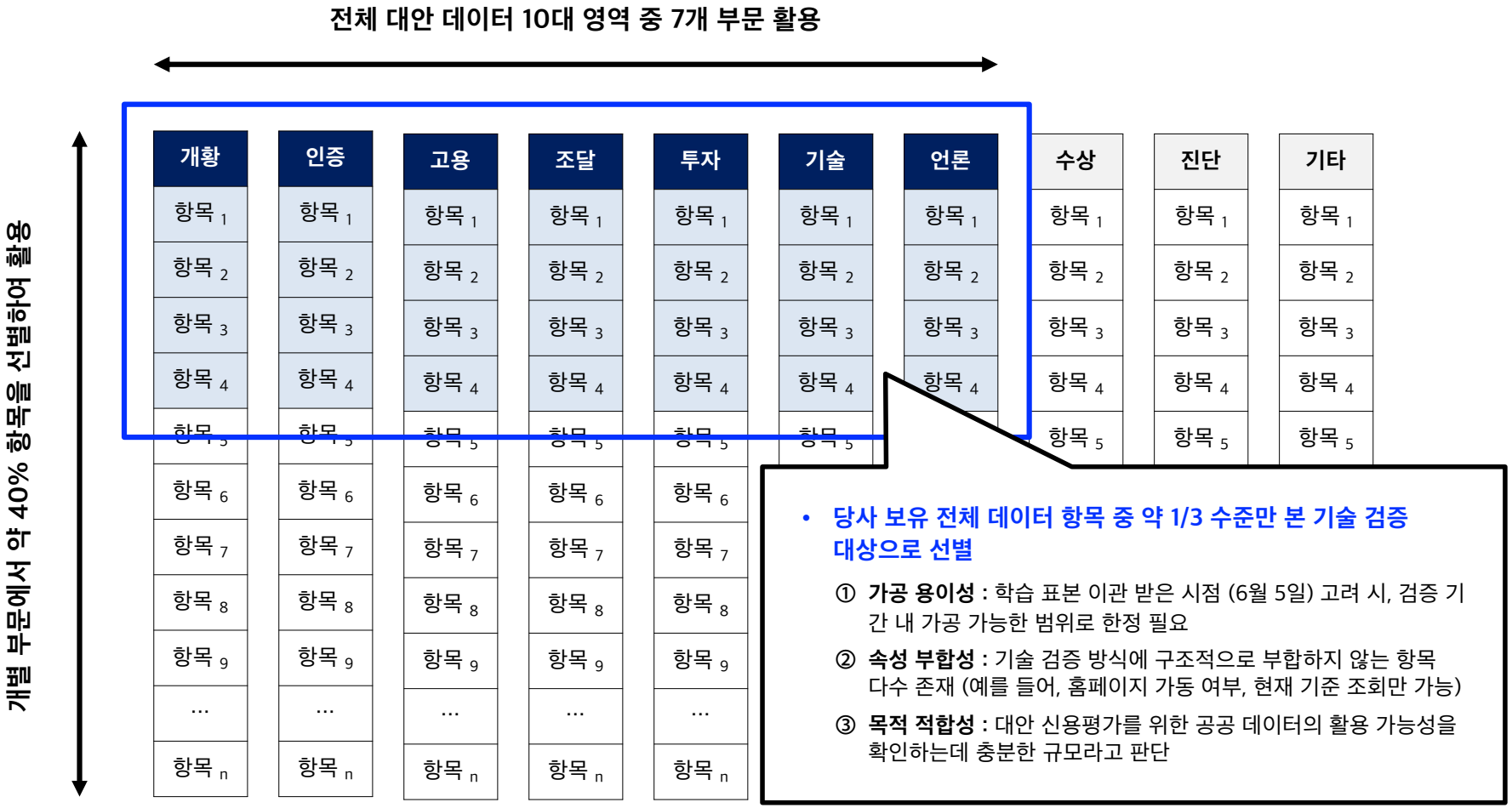
- 최종적으로 변환된 공공 데이터 항목 별, 실제 부도 표본과 정상 표본간의 유의미한 통계적 차이가 존재하는지 분석하고 결과 표현

➢ 성김화 (Coarse Classing) 등을 통해 양 집단 간의 차이를 직관적으로 표현

최종적으로 약 50개 유의미한 지표 도출

가 데이터 선별

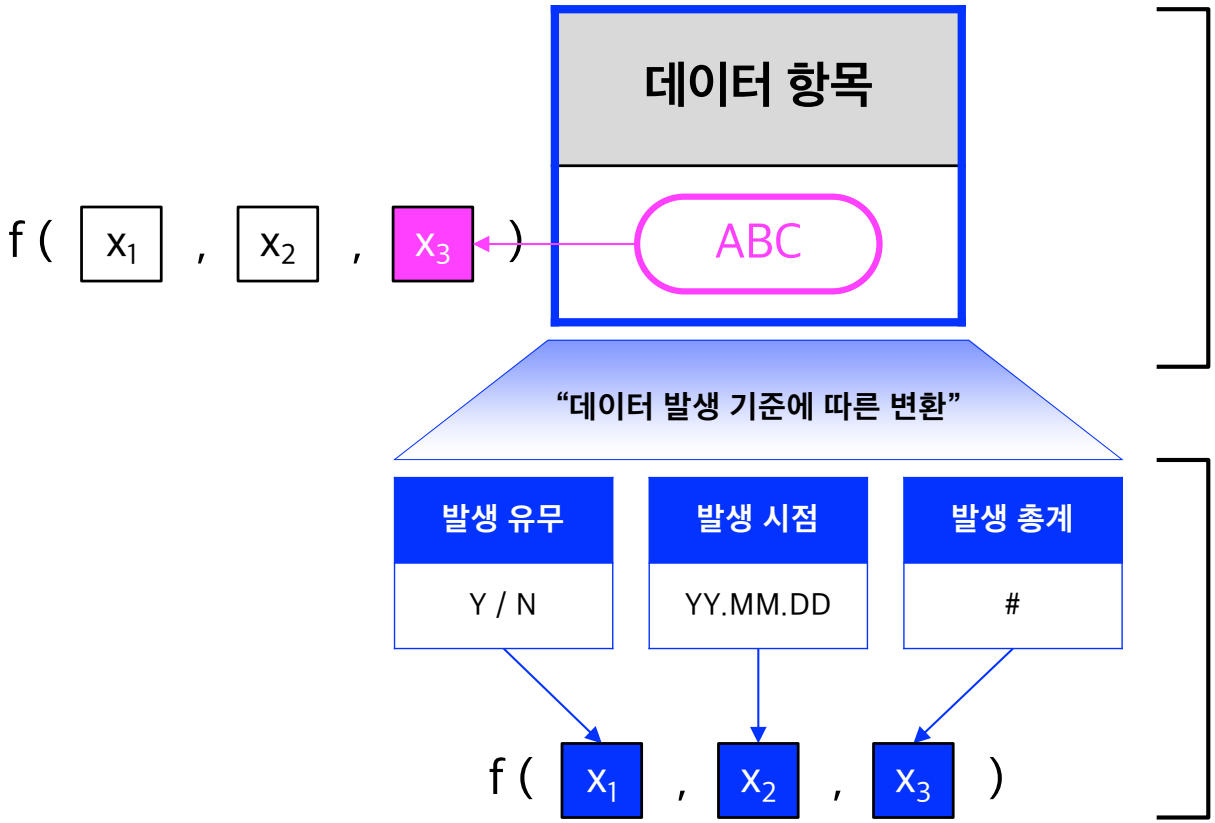
데이터 선별 과정에서는 시간 제약에 따른 가공 용이성, 시계열 분석 부합성, 마지막으로 기술 검증의 목적 부합성에 따라 전체 보유 데이터 중 30% 수준의 항목을 선별하였습니다.



나 데이터 가공

전처리 단계에서는 공공 데이터의 세부 내용 외 데이터 발생 로그를 추출하여 발생 시점, 발생 채널, 발생 규모 등의 파생 데이터를 표준화 된 형태로 가공하였습니다.

데이터 활용 / 가공 방법론 유형화



기존 데이터 활용 방식

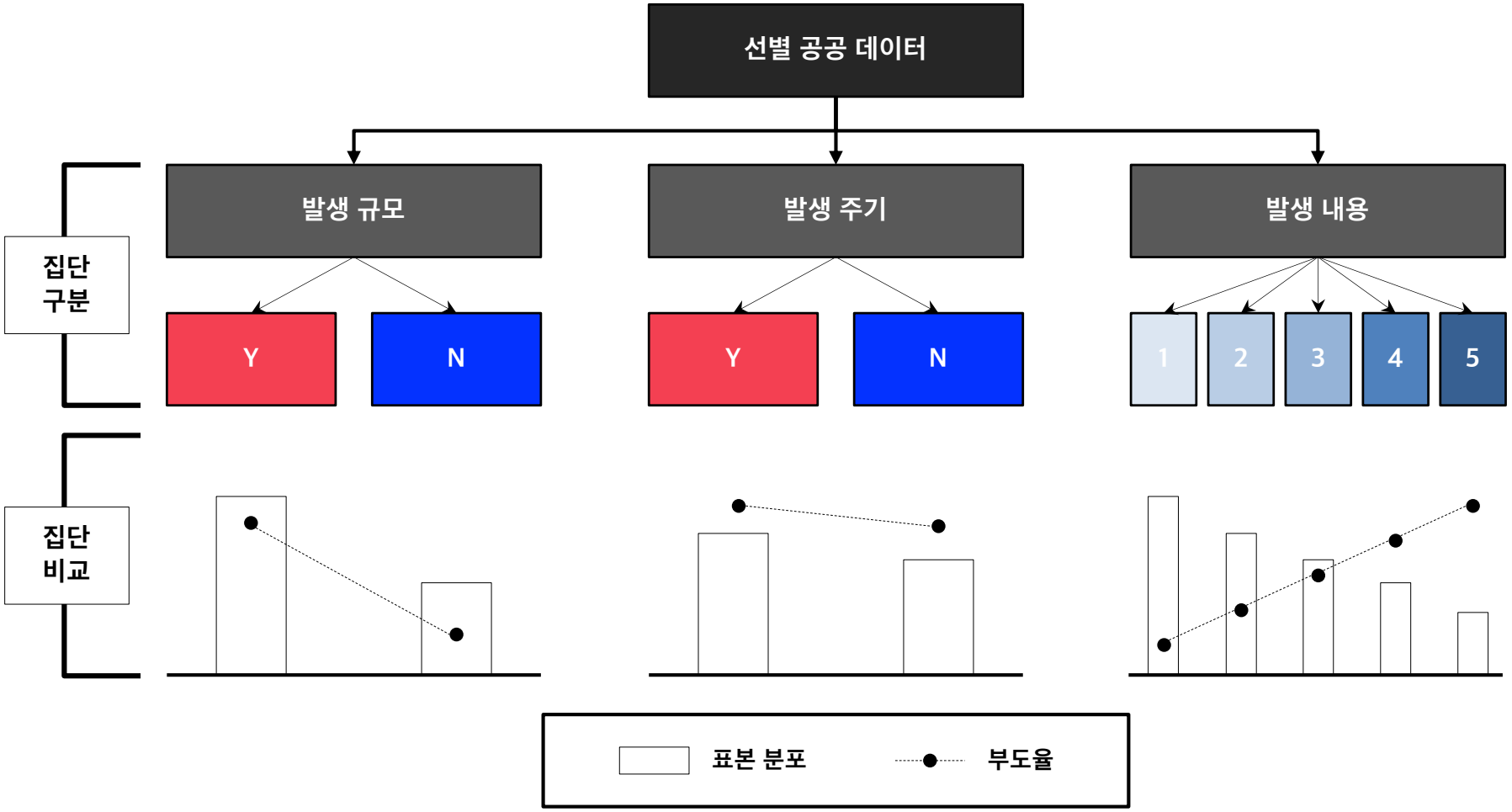
- **구체적인 데이터 값을 분석의 대상으로 설정하여 모형 개발**
  - ➢ 예시) 21년 매출 453억 원에서, '453' 활용
  - ➢ 예시) 21년 12월 고용 20명에서, '20' 활용

공공 데이터 활용 방식

- **세부적인 데이터 값 보다는, 데이터 항목 발생과 관련된 로그 활용**
  - ➢ 예시) 21년 매출 453억 원에서, 매출 발생 시점 '21년' 활용
  - ➢ 예시) 21년 12월 고용 20명에서, 데이터 존재하므로 'Y' 활용

다 데이터 분석

데이터 분석을 위해서 공공 데이터의 발생 여부, 규모, 주기, 그리고 세부 내용을 이원화 혹은 범주화 된 집단으로 구부하여 유의미한 부도율의 차이가 존재하는지 분석하였습니다.



데이터 유의성 분석 결과 (1/10)

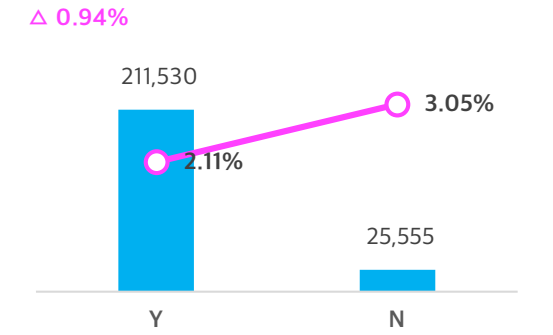
규모 | 시점 | 내용

세부 데이터 항목 별 집단 간 유의미한 부도율의 차이가 발생하는 지표는 약 50개 수준이었으며, 세부 내용은 이하와 같습니다.

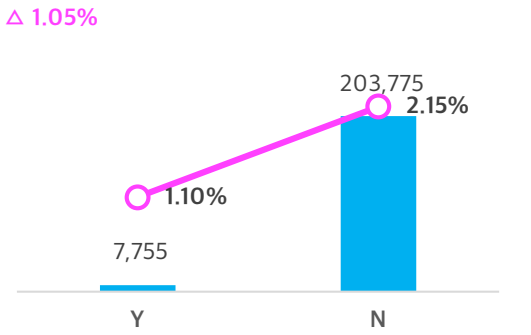
[데이터 발생 집단 비교 분석]

- **데이터 발생 여부 / 규모 / 빈도를 이항 형태로 구분하고, 이에 따른 집단 간 부도율 비교**
  - ➢ 전체 제공 표본 평균 부도율 **2.21%**
  - ➢ 분석 대상 표본 평균 부도율 **2.11%**

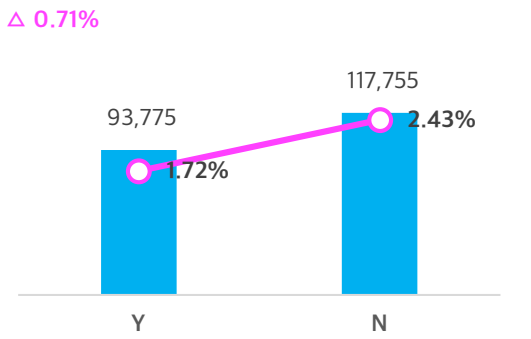
01. 식별 데이터 1 발생 여부



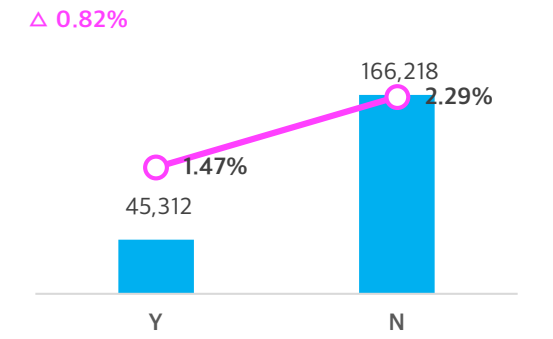
02. 개황 데이터 2 발생 여부



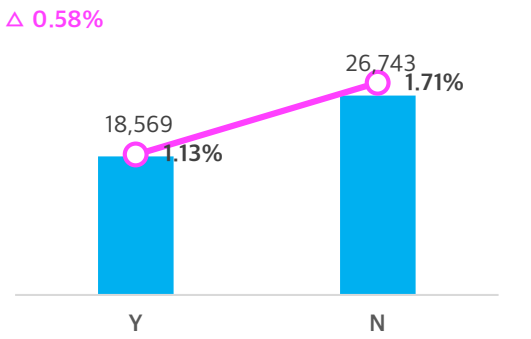
03. 개황 데이터 3 발생 여부



04. 개황 데이터 4 발생 여부



05. 개황 데이터 5 여부



데이터 유의성 분석 결과 (2/10)

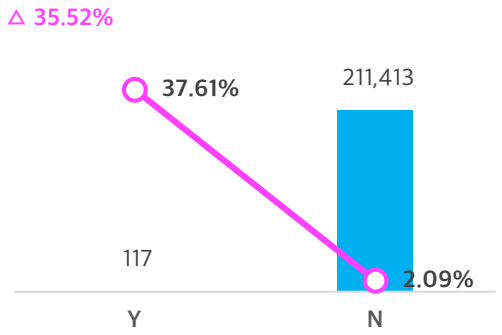
규모 | 시점 | 내용

세부 데이터 항목 별 집단 간 유의미한 부도율의 차이가 발생하는 지표는 약 50개 수준이었으며, 세부 내용은 이하와 같습니다.

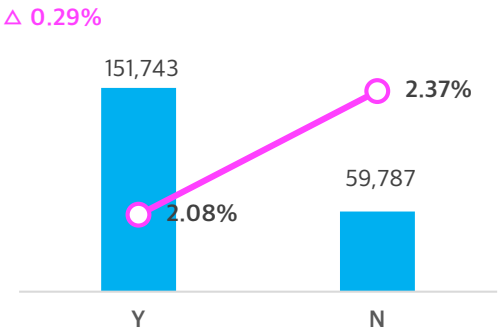
[데이터 발생 집단 비교 분석]

- **데이터 발생 여부 / 규모 / 빈도를 이항 형태로 구분하고, 이에 따른 집단 간 부도율 비교**
  - ➢ 전체 제공 표본 평균 부도율 **2.21%**
  - ➢ 분석 대상 표본 평균 부도율 **2.11%**

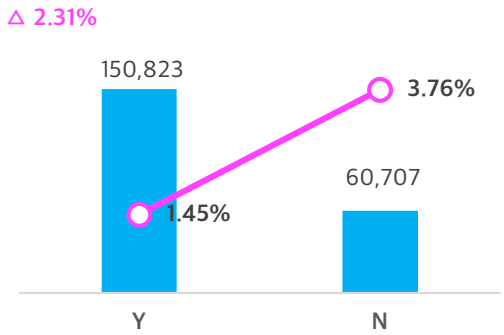
06. 세금 데이터 발생 여부



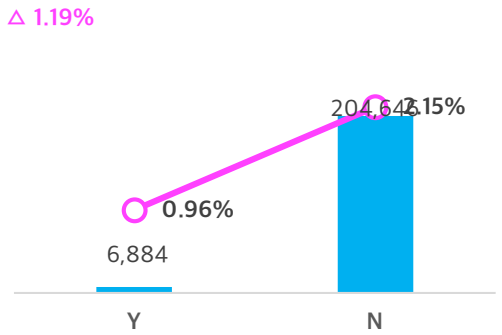
07. 지역 유형 데이터 발생 여부



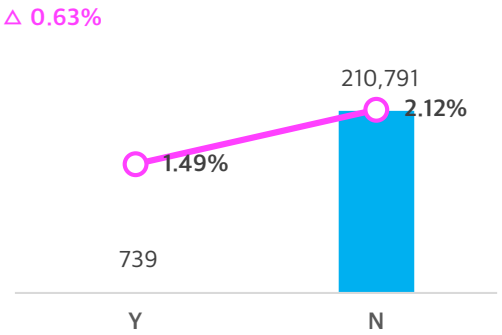
08. 공시 데이터 발생 여부



09. 대표 구성 데이터 발생 여부



10. 인증 데이터 1 발생 여부



데이터 유의성 분석 결과 (3/10)

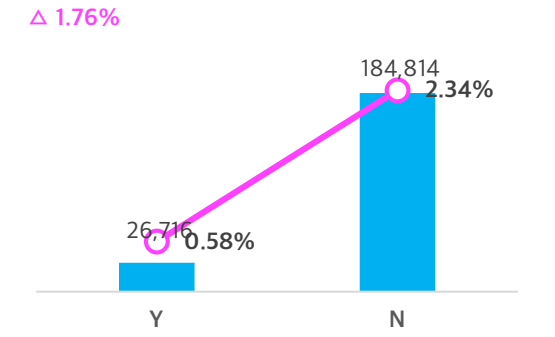
규모 | 시점 | 내용

세부 데이터 항목 별 집단 간 유의미한 부도율의 차이가 발생하는 지표는 약 50개 수준이었으며, 세부 내용은 이하와 같습니다.

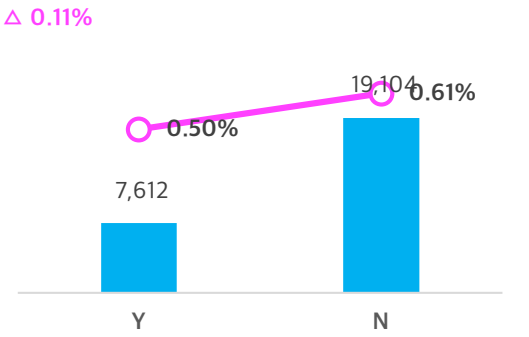
[데이터 발생 집단 비교 분석]

- **데이터 발생 여부 / 규모 / 빈도를 이항 형태로 구분하고, 이에 따른 집단 간 부도율 비교**
  - ➢ 전체 제공 표본 평균 부도율 **2.21%**
  - ➢ 분석 대상 표본 평균 부도율 **2.11%**

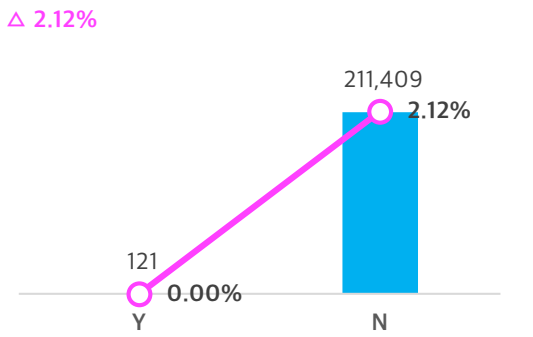
11. 인증 데이터 2 발생 여부



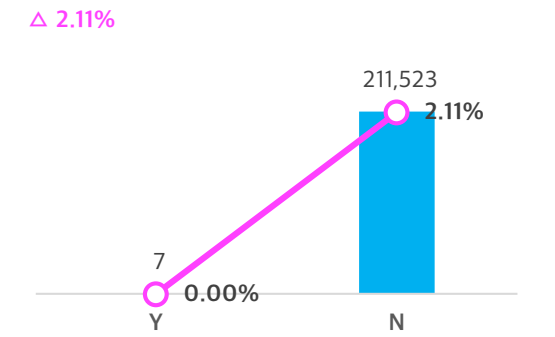
12. 인증 데이터 2 갱신 발생 여부



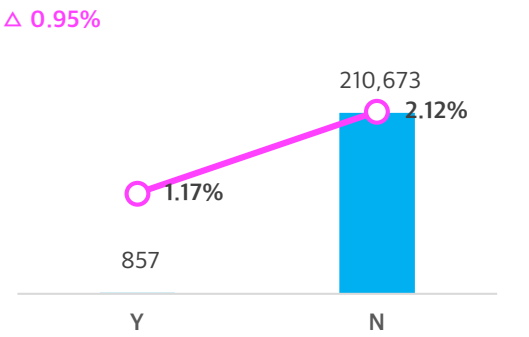
13. 인증 3 데이터 발생 여부



14. 인증 4 데이터 발생 여부



15. 수출입 데이터 1 발생 여부



데이터 유의성 분석 결과 (4/10)

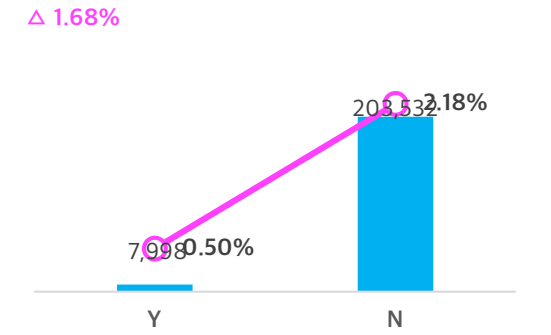
규모 | 시점 | 내용

세부 데이터 항목 별 집단 간 유의미한 부도율의 차이가 발생하는 지표는 약 50개 수준이었으며, 세부 내용은 이하와 같습니다.

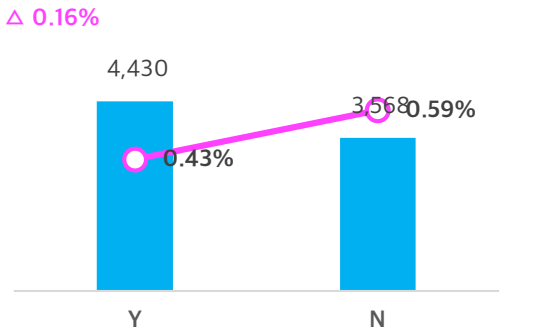
[데이터 발생 집단 비교 분석]

- **데이터 발생 여부 / 규모 / 빈도를 이항 형태로 구분하고, 이에 따른 집단 간 부도율 비교**
  - ➢ 전체 제공 표본 평균 부도율 **2.21%**
  - ➢ 분석 대상 표본 평균 부도율 **2.11%**

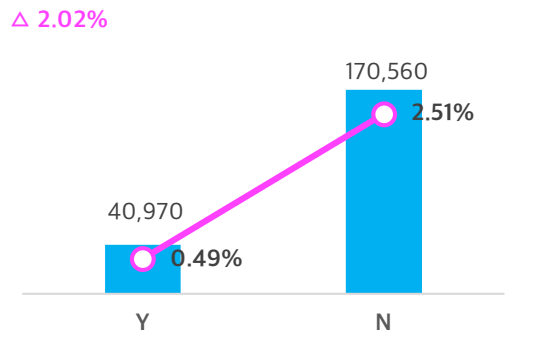
16. 조달 데이터 1 발생 여부



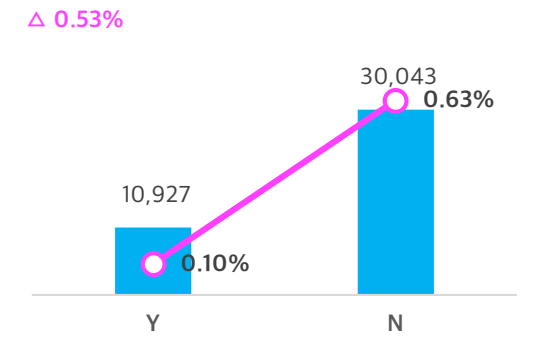
17. 조달 데이터 1 복수 발생 여부



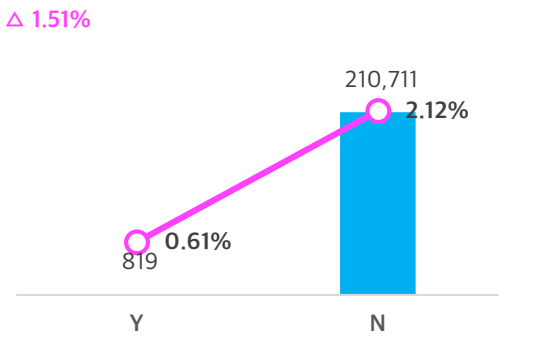
18. 기술 데이터 1 발생 여부



19. 기술 데이터 1 복수 발생 여부



20. 인증 데이터 5 발생 여부



데이터 유의성 분석 결과 (5/10)

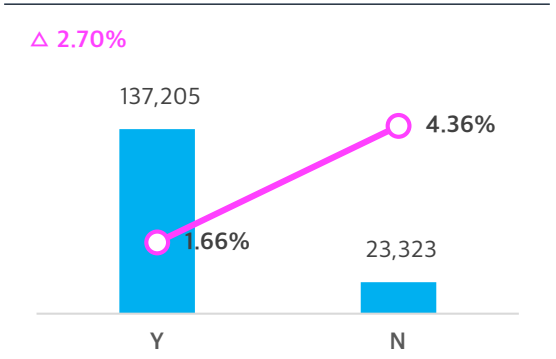
규모 | 시점 | 내용

세부 데이터 항목 별 집단 간 유의미한 부도율의 차이가 발생하는 지표는 약 50개 수준이었으며, 세부 내용은 이하와 같습니다.

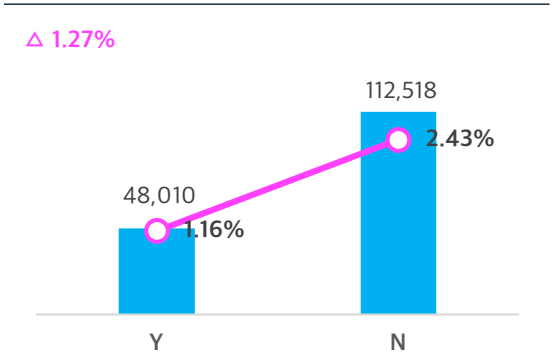
[데이터 발생 집단 비교 분석]

- **데이터 발생 여부 / 규모 / 빈도를 이항 형태로 구분하고, 이에 따른 집단 간 부도율 비교**
  - ➢ 전체 제공 표본 평균 부도율 **2.21%**
  - ➢ 분석 대상 표본 평균 부도율 **2.11%**

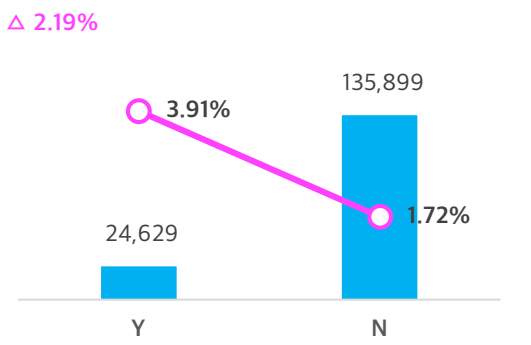
21. 고용 데이터 1 발생 여부



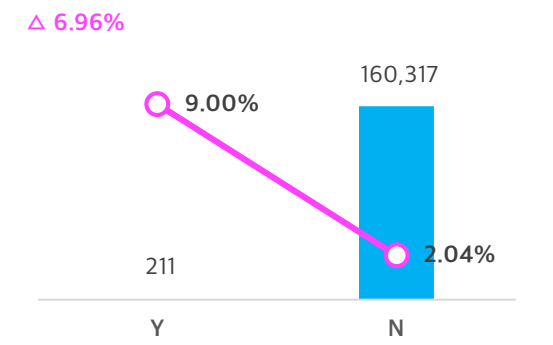
22. 고용 데이터 1 연속 발생 여부



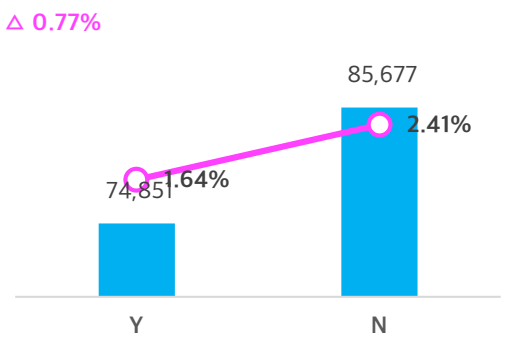
23. 고용 데이터 1 변화 여부



24. 고용 데이터 2 발생 여부



25. 고용 데이터 3 차이 여부



데이터 유의성 분석 결과 (6/10)

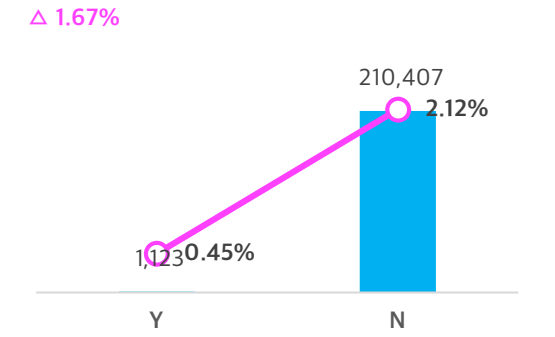
규모 | 시점 | 내용

세부 데이터 항목 별 집단 간 유의미한 부도율의 차이가 발생하는 지표는 약 50개 수준이었으며, 세부 내용은 이하와 같습니다.

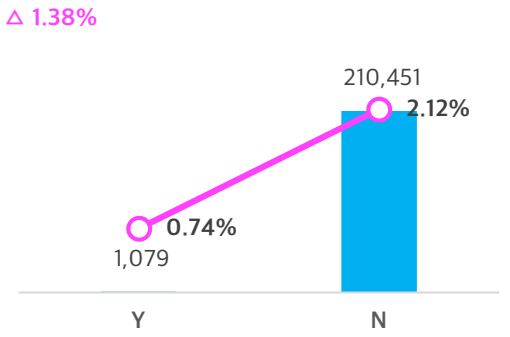
[데이터 발생 집단 비교 분석]

- **데이터 발생 여부 / 규모 / 빈도를 이항 형태로 구분하고, 이에 따른 집단 간 부도율 비교**
  - ➢ 전체 제공 표본 평균 부도율 **2.21%**
  - ➢ 분석 대상 표본 평균 부도율 **2.11%**

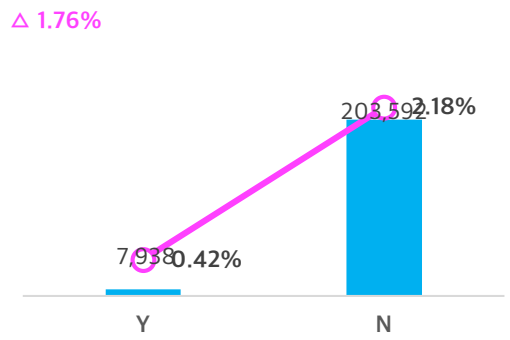
26. 수출입 데이터 2 발생 여부



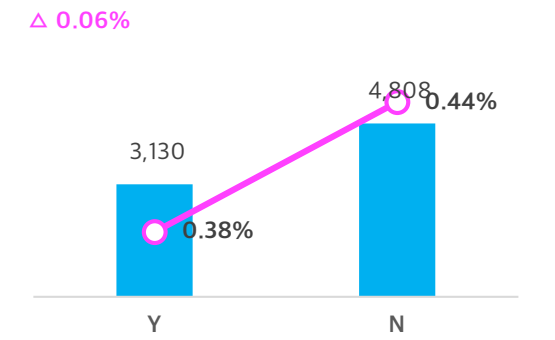
27. 투자 데이터 발생 여부



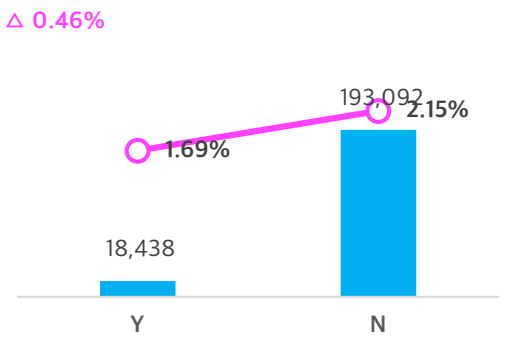
28. 기술 데이터 2 발생 여부



29. 기술 데이터 2 복수 발생 여부



30. IP 데이터 1 발생 여부



데이터 유의성 분석 결과 (7/10)

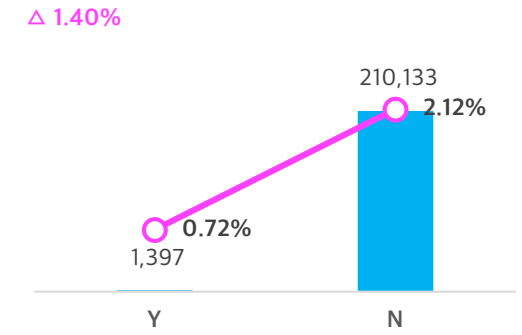
규모 | 시점 | 내용

세부 데이터 항목 별 집단 간 유의미한 부도율의 차이가 발생하는 지표는 약 50개 수준이었으며, 세부 내용은 이하와 같습니다.

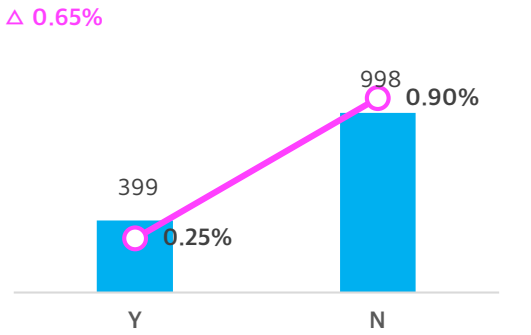
[데이터 발생 집단 비교 분석]

- **데이터 발생 여부 / 규모 / 빈도를 이항 형태로 구분하고, 이에 따른 집단 간 부도율 비교**
  - ➢ 전체 제공 표본 평균 부도율 **2.21%**
  - ➢ 분석 대상 표본 평균 부도율 **2.11%**

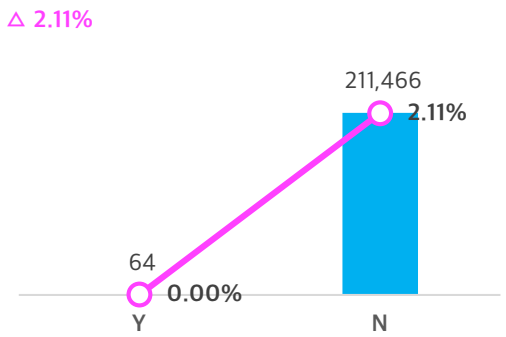
31. IP 데이터 2 발생 여부



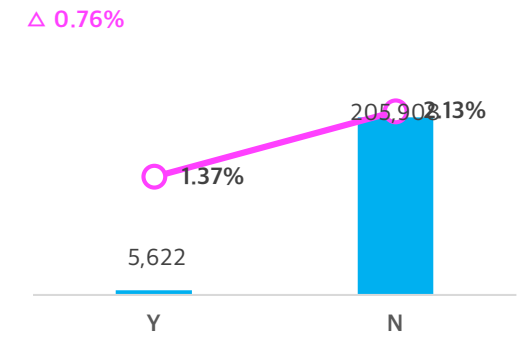
32. IP 데이터 2 복수 발생 여부



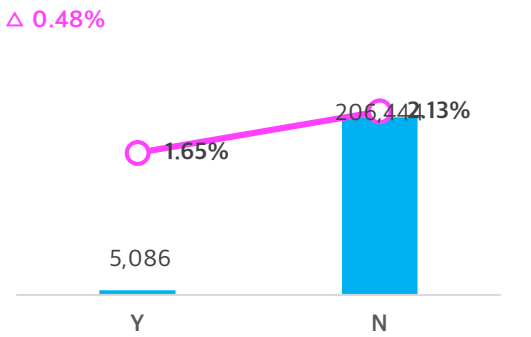
33. IP 데이터 3 발생 여부



34. IP 데이터 3 복수 발생 여부



35. IP 데이터 4 발생 여부



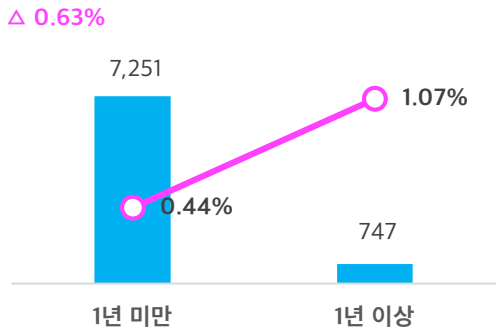
데이터 유의성 분석 결과 (8/10)

세부 데이터 항목 별 집단 간 유의미한 부도율의 차이가 발생하는 지표는 약 50개 수준이었으며, 세부 내용은 이하와 같습니다.

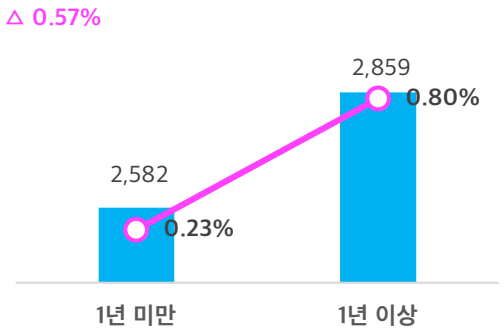
[데이터 발생 집단 비교 분석]

- **데이터 발생 여부 / 규모 / 빈도를 이항 형태로 구분하고, 이에 따른 집단 간 부도율 비교**
  - ➢ 전체 제공 표본 평균 부도율 **2.21%**
  - ➢ 분석 대상 표본 평균 부도율 **2.11%**

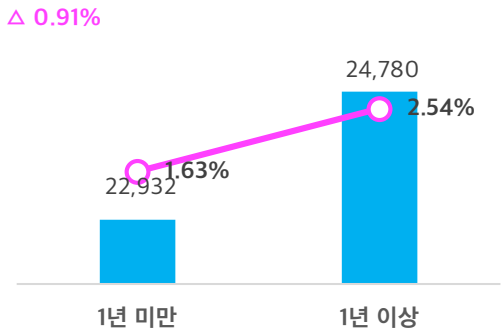
36. 조달 데이터 1 발생 경과 기간



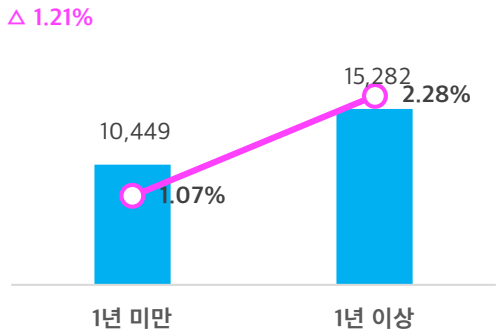
37. 기술 데이터 2 발생 경과 기간



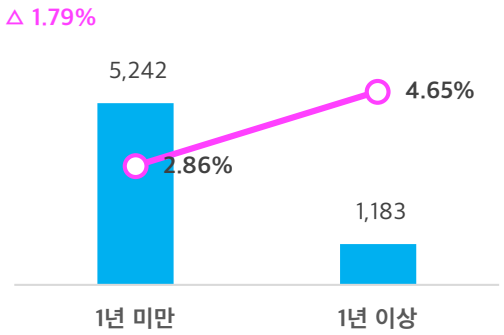
38. IP 데이터 5 발생 경과 기간



39. IP 데이터 6 발생 경과 기간



40. 뉴스 데이터 발생 경과 기간



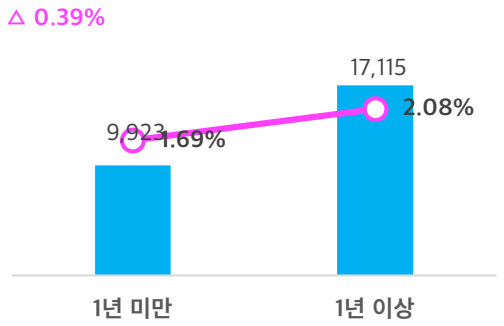
데이터 유의성 분석 결과 (9/10)

세부 데이터 항목 별 집단 간 유의미한 부도율의 차이가 발생하는 지표는 약 50개 수준이었으며, 세부 내용은 이하와 같습니다.

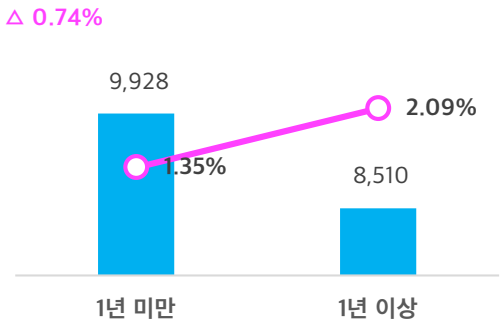
[데이터 발생 집단 비교 분석]

- **데이터 발생 여부 / 규모 / 빈도를 이항 형태로 구분하고, 이에 따른 집단 간 부도율 비교**
  - ➢ 전체 제공 표본 평균 부도율 **2.21%**
  - ➢ 분석 대상 표본 평균 부도율 **2.11%**

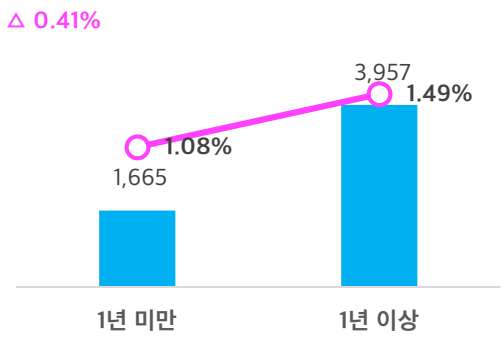
41. IP 데이터 1 발생 경과 기간



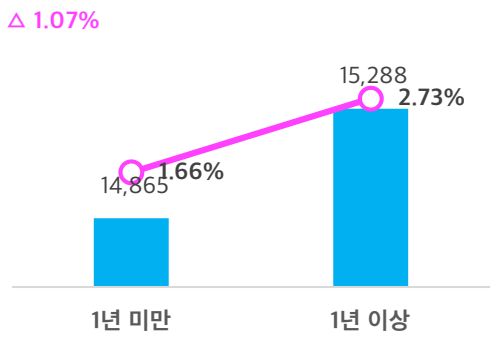
42. IP 데이터 7 발생 경과 기간



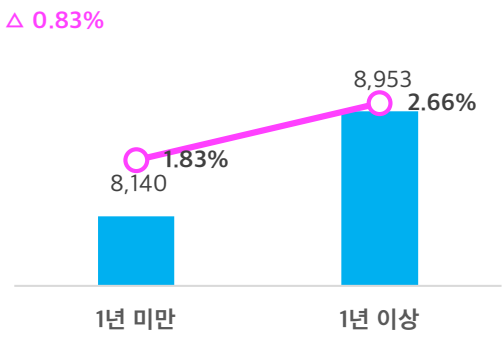
43. IP 데이터 8 발생 경과 기간



44. IP 데이터 9 발생 경과 기간



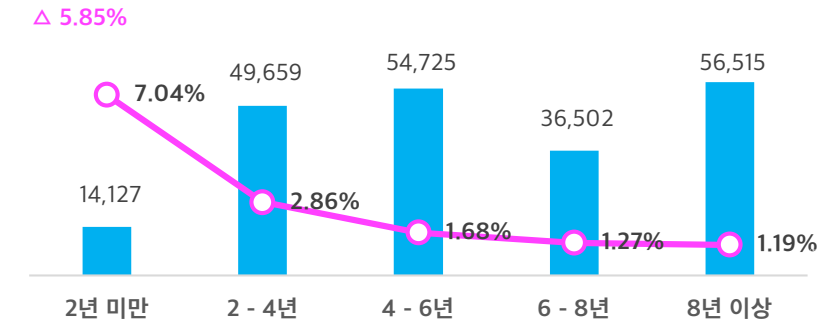
45. IP 데이터 10 발생 경과 기간



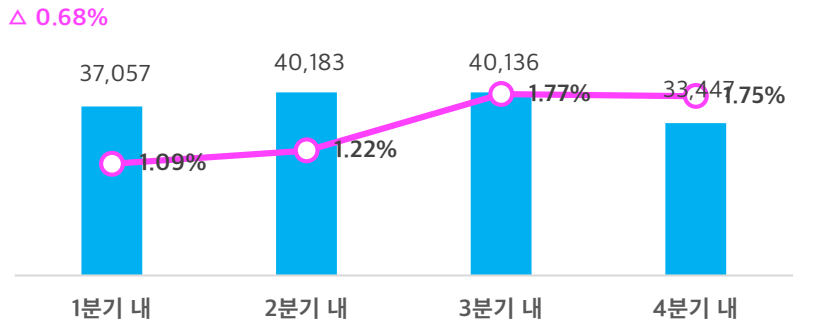
데이터 유의성 분석 결과 (10/10)

세부 데이터 항목 별 집단 간 유의미한 부도율의 차이가 발생하는 지표는 약 50개 수준이었으며, 세부 내용은 이하와 같습니다.

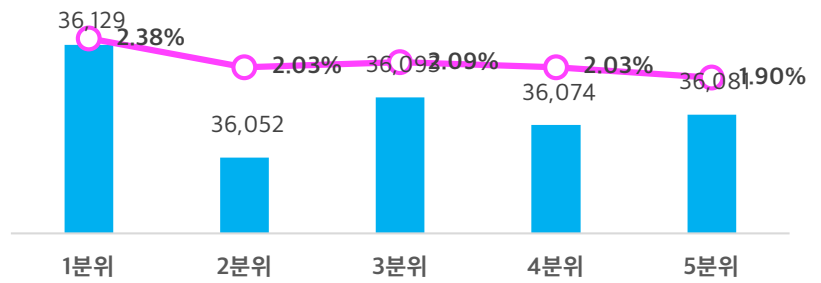
46. 공공 데이터 발생 기간



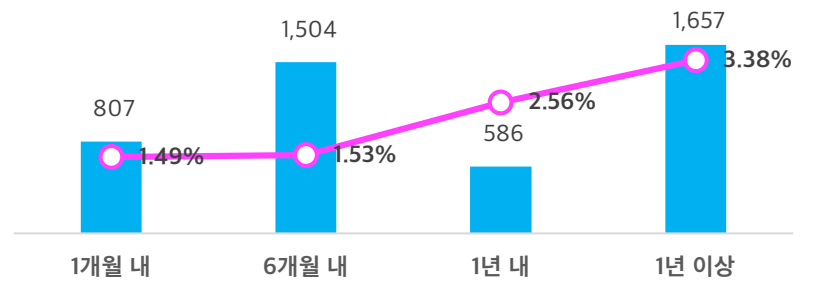
47. 공시 데이터 발생 경과 기간



48. 주소 데이터 가치 등위 구분



49. 뉴스 데이터 발생 경과 기간

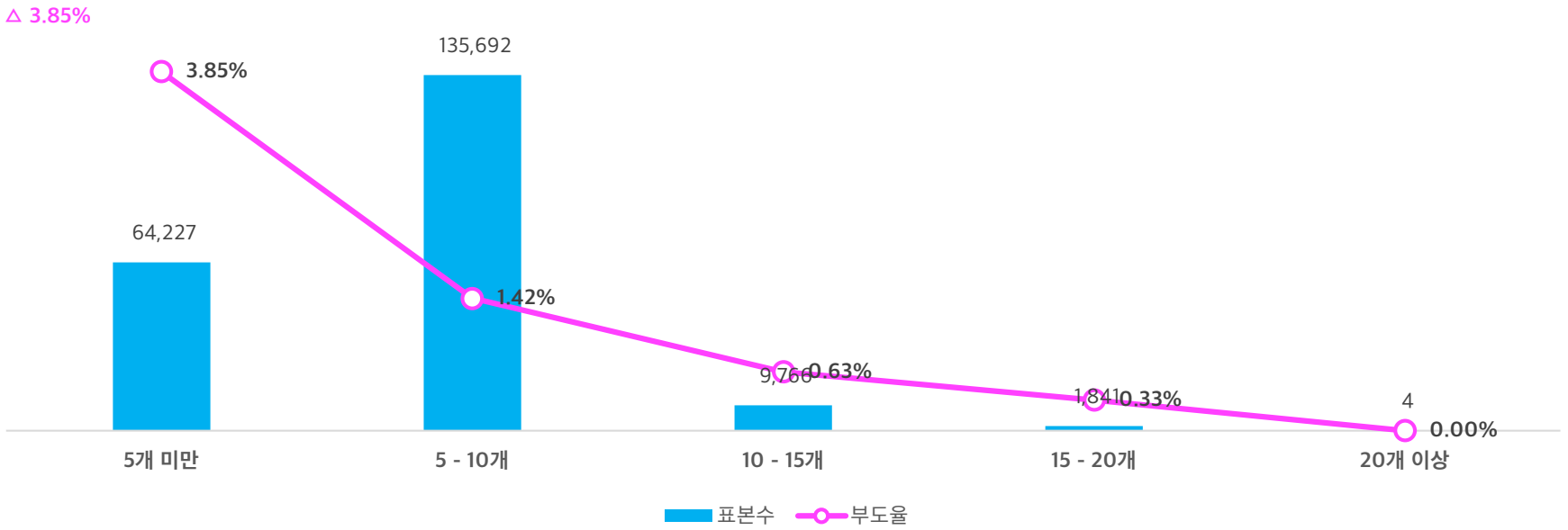


데이터 유의성 분석 결론

규모 | 시점 | 내용

최종적으로, 공공 데이터 발생 규모에 따른 부도율의 차이는 집단 별로 극명한 형태로 존재하였으며, 이에 따라 유의미한 상관관계가 존재함을 의미합니다.

50. 공공데이터 발생 규모에 따른 부도율 비교



	5개 미만	5 - 10개	10 - 15개	15 - 20개	20개 이상
표본수	64,227	135,692	9,766	1,841	4
부도율	3.85%	1.42%	0.63%	0.33%	0.00%

Agenda

기술 검증 추진 배경

공공 데이터 유의성 검증

신용평가 (예측) 모형 검증

현장 테스트 결과

부도 예측력 평가 개요

데이터 유의성의 다음 단계로 실제 공공 데이터를 활용한 부도율 예측 모델 개발에 대한 작업을 진행하였으며 세부 프로세스는 이하와 같습니다.

AI 부도 예측 모형 개발 프로세스

변수 선정

- 전체 모집단 (211,530 개) 대상으로 적용 가능한 독립변수 중, 데이터 유의성 검사에서 부도 및 정상 표본간 부도율 1% 이상 차이나는 30개 항목 선정

알고리즘 채택

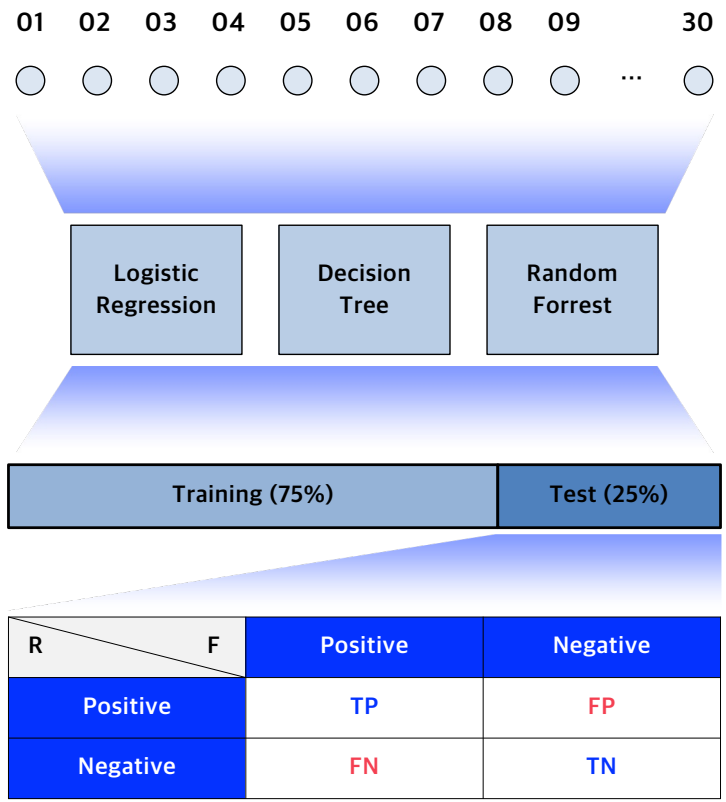
- 가장 대표적인 인공지능 분석 방법인 Logistic Regression, Decision Tree, Random Forest 세 가지 방식으로 분석 진행

표본 구성

- 학습 및 평가 표본 비율을 7.5 : 2.5 형태로 구성하고, 계층화 Sampling 활용하여 각각 158,647개, 52,883개로 구분

성능 평가

- Confusion Matrix (혼돈행렬) 방식을 통해 성능 평가 진행하며, 이를 통해 Accuracy, Precision, Recall, F1, ROC_AUC 지표 도출



R \ F	Positive	Negative
Positive	TP	FP
Negative	FN	TN

부도 예측력 평가 결과 (1/5)

주요 인공지능 모형에 따른 분석을 완료하였으며 혼돈행렬 (Confusion Matrix) 상의 결과는 아래와 같고, 이중 Random Forest 방식이 가능 높은 성능을 기록하였습니다.

Logistic Regression

		예측		
		정상	부도	총계
실제	정상	**51,801**	**3**	51,804
	부도	**1,077**	**2**	1,079
	총계	52,878	5	52,883

모형 성능 지표	점수
Accuracy	97.96%
Precision	40.00%
Recall	0.19%
F1 Score	0.37%
ROC_AUC	50.09%

Decision Tree

		예측		
		정상	부도	총계
실제	정상	51,493	311	51,804
	부도	393	686	1,079
	총계	51,886	997	52,883

모형 성능 지표	점수
Accuracy	98.69%
Precision	68.81%
Recall	63.58%
F1 Score	66.09%
ROC_AUC	81.49%

Random Forest

		예측		
		정상	부도	총계
실제	정상	51,582	222	51,804
	부도	402	677	1,079
	총계	51,984	899	52,883

모형 성능 지표	점수
Accuracy	98.80%
Precision	75.31%
Recall	62.74%
F1 Score	68.45%
ROC_AUC	81.16%

채택

부도 예측력 평가 결과 (2/5)

보다 구체적으로 살펴보면, Random Forest 분석의 Accuracy, Precision, Recall 지표는 각각 98%, 75%, 63% 수준이었고, F1 Score 및 ROC_AUC 지표도 각각 68%, 81% 수준으로 양호합니다.

Confusion Matrix (혼돈행렬) 구조

		모형 예측	
		정상 (N)	부도 (P)
실제 발생	정상 (N)	TN	FP
	부도 (P)	FN	TP

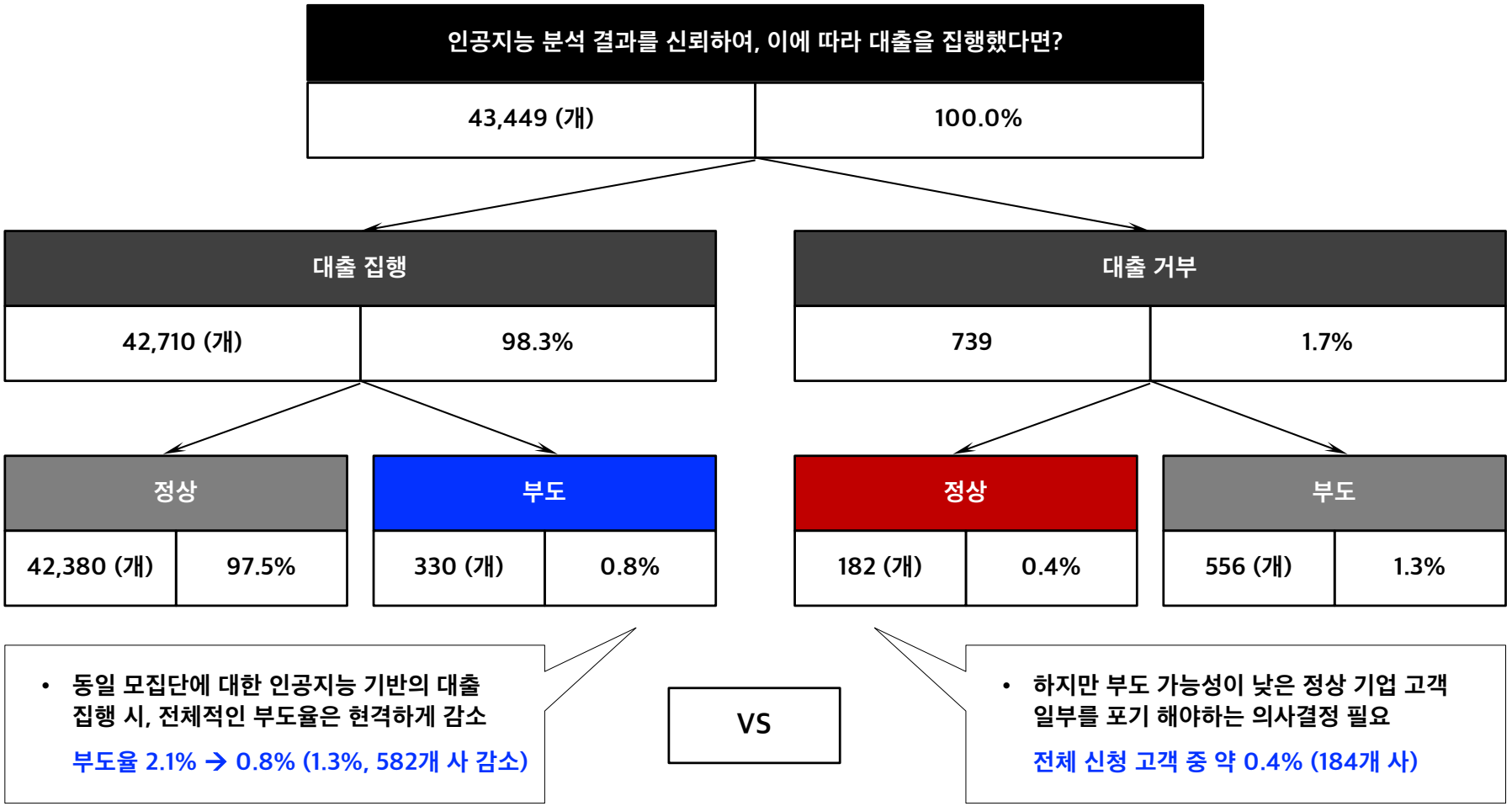
- **TN** 정상 기업을 정상으로 예측한 경우
- **FP** 정상 기업을 부도로 오측한 경우 **(1종 오류)**
- **TP** 부도 기업을 부도로 예측한 경우
- **FN** 부도 기업을 정상으로 예측한 경우 **(2종 오류)**

인공지능 주요 성능 평가 기준 설명

모형 성능 지표	산출 방식	성능 기준
Accuracy	**전체 표본 중에서 실제 정상 / 부도 기업을 정확하게 예측한 비율 : 98.80%** [공식] (TN+TP) / (TN + TP + FN + FP)	**100% 근접할수록 훌륭한 모형** (세부적인 기준은 주관적인 판단 필요)
Precision	**부도라고 예측한 기업 중에서 실제로 부도가 발생한 표본 비중 : 75.31%** [공식] TP / (TP + FP)	
Recall	**실제 부도가 발생 기업 중에서 부도라고 정확하게 예측한 비율 : 62.74%** [공식] TP / (TP + FN)	
F1 Score	N/A	미흡 : 0 - 50% 사용 가능 : 50 - 80% → **68.45%** 우수 : 80 - 90% 매우 우수 : 90 - 100%
ROC_AUC	N/A	미흡 : 0 - 70% 사용 가능 : 70 - 80% 우수 : 80 - 90% → **81.16%** 매우 우수 : 90 - 100%

부도 예측력 평가 결과 (3/5)

실제 해당 모형에 따른 대출 의사결정을 실행할 경우, 기존 표본에서 부도 차주를 582개를 감축할 수 있으며 부도율이 2.1% 수준에서 1.3% 하락한 0.8% 수준으로 유지할 수 있습니다.



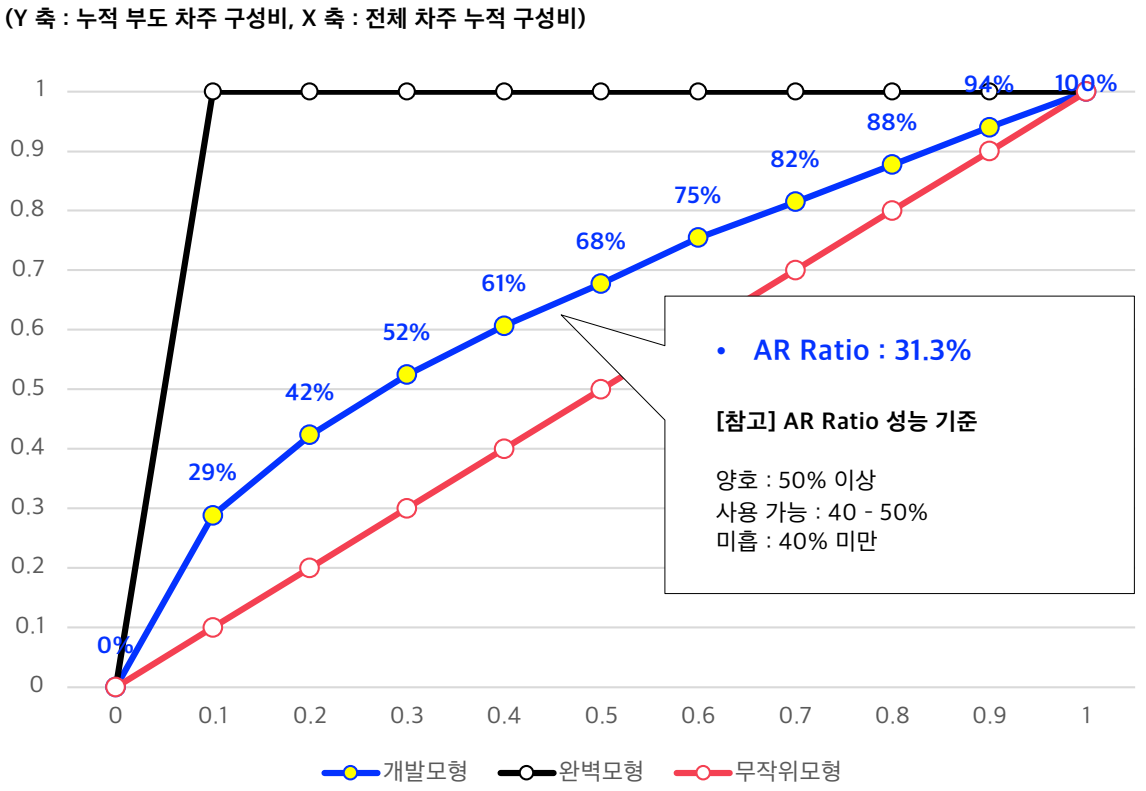
부도 예측력 평가 결과 (4/5)

하지만 신용평가 모형에 대한 성능 평가 방식인 AR (Accuracy Ratio) 수준은 약 31% 수준으로, 이에 대한 보완이 필요합니다.

신용등급 분포표 (일괄 10분위)

분위	표본	부도 여부		부도율
		Y	N	
10	5,290	334	4,956	6.31%
9	5,287	156	5,131	2.95%
8	5,288	117	5,171	2.21%
7	5,290	95	5,195	1.80%
6	5,288	82	5,206	1.55%
5	5,287	89	5,198	1.68%
4	5,289	70	5,219	1.32%
3	5,287	72	5,215	1.36%
2	5,288	73	5,215	1.38%
1	5,289	69	5,220	1.30%
합계	**52,883**	**1,157**	**51,726**	**2.19%**

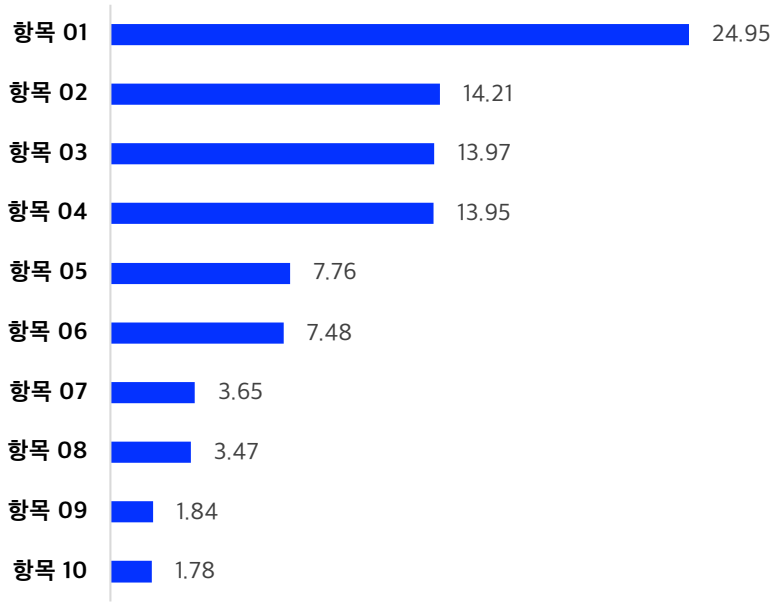
Accuracy Ratio (AR) 기반 등급체계 평가



부도 예측력 평가 결과 (5/5)

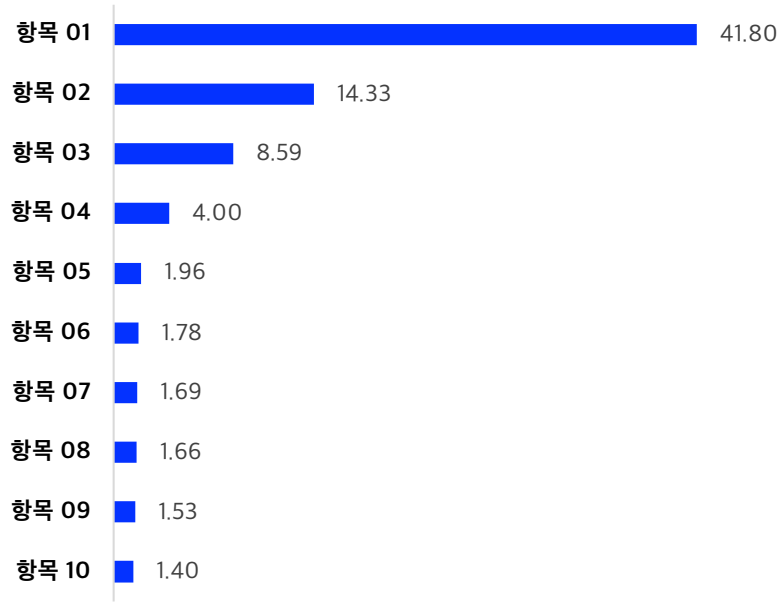
AR 성능 평가의 저조 및 1종 / 2종 오류 발생 원인은 인공지능 모형의 특성상 편중된 가중치에 의거한 현상이라고 판단합니다.

최초 시도 (2020) AI 가중치 부여



- **전체 활용 독립 변수 47개 중, TOP 10 가중치 차지 비중이 93% 이상이며, 이하 변수의 경우 모두 1% 이하로 전체 모형에서 수행하는 역할 미미**

금번 시도 (2022) AI 가중치 부여



- **전체 활용 독립 변수 약 30개 중, TOP 10 가중치 차지 비중이 78% 이상이며, 여전히 변수 의존도가 높아, 기타 데이터에 대한 활용도가 제약**

PULSE SCORING MODEL (PSM)

당사는 이에 대한 문제 인식으로 장기적인 관점에서 개별 공공 데이터에 대한 미세한 신호 포착과 이를 기업 리스크 예측 모형에 반영한 PULSE 알고리즘을 개발하고 있습니다.

PULSE 알고리즘 소개

- **공공 데이터에 대한 관점**
  - 귀납적인 관찰 결과, 건강한 기업일수록 사업 수행 경과에 따른 공공 데이터가 규칙적이며 강하게 발생
  - 반대로, 부실 기업일수록 공공 데이터가 제한적이며 불규칙적인 형태로 발생하는 상관관계 존재
- **신용평가 연계 방법론**
  - 공공 데이터는 개별 기업의 건강 상태를 직간접적으로 추론할 수 있는 맥박과 같은 신호라고 간주
  - 따라서, '미세한' 신호를 포착하고 이를 신용평가에 반영할 수 있는 Scoring System 개발 연구

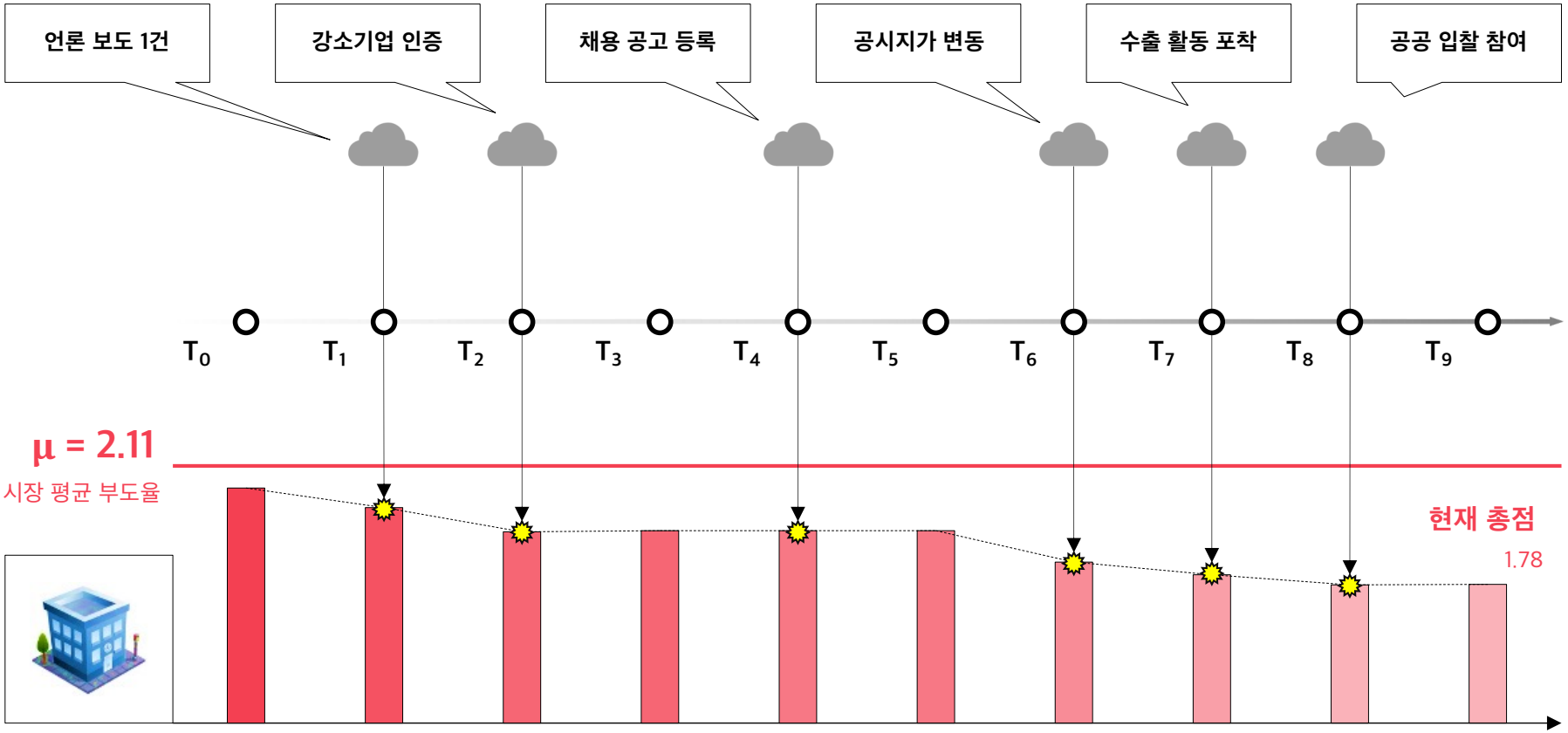
➢ **공공데이터가 발생할 때마다 개별 기업의 리스크를 세부적으로 조정하며, '미세한' 차이가 희석되지 않도록 등급화**

The PULSE

PULSE SCORING MODEL (PSM) (1/3)

PULSE 알고르짐은 특정 데이터 유형에 대한 과도한 가중치 반영을 지양하고, 개별 공공 데이터 발생 현상에 따른 점수 부여를 실시함으로써 리스크 등급을 책정합니다.

공공데이터 발생에 따른 기업 부도 가능성 조정



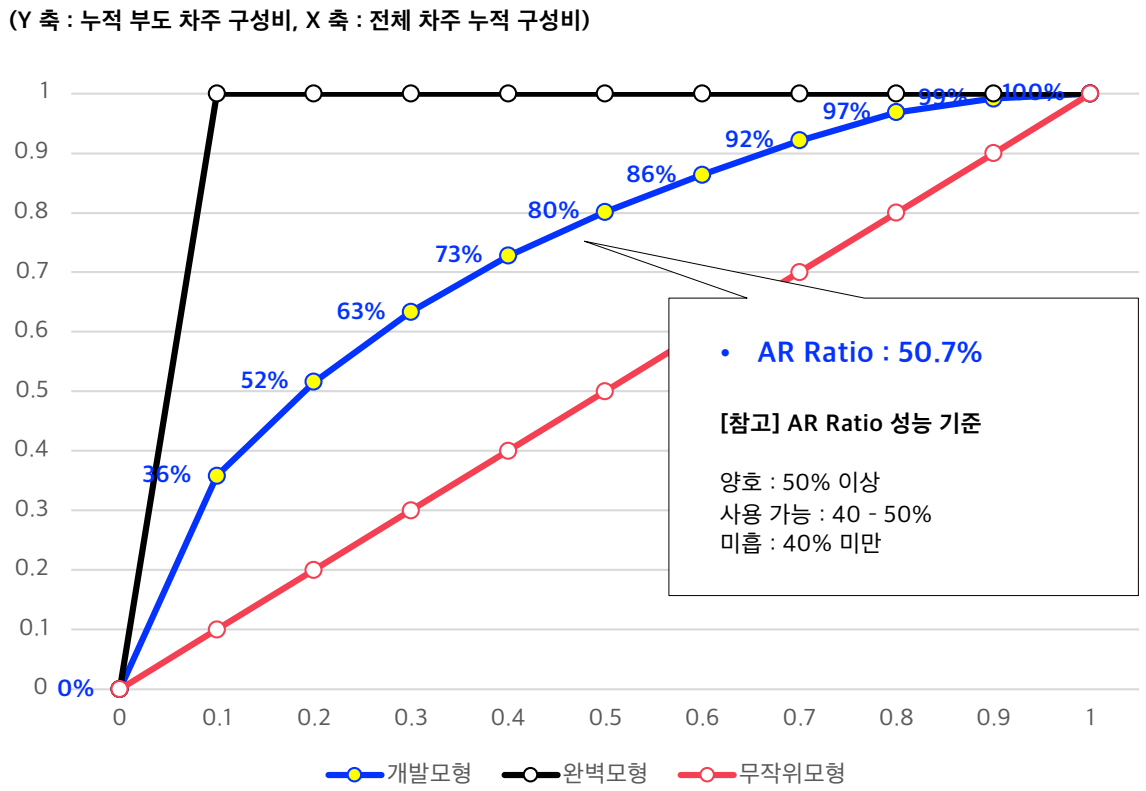
PULSE SCORING MODEL (PSM) (2/3)

PULSE 알고리즘에 기반한 점수 등급에 따른 분위 구분은 이하와 같으며, 이에 따른 AR 점수는 50% 이상으로 크게 향상되었습니다.

신용등급 분포표 (일괄 10분위)

분위	표본	부도 여부		부도율
		Y	N	
10	21,151	1,602	19,549	7.57%
9	21,148	705	20,443	3.33%
8	21,126	526	20,600	2.49%
7	21,180	422	20,758	1.99%
6	21,147	330	20,817	1.56%
5	21,142	277	20,865	1.31%
4	21,174	258	20,916	1.22%
3	21,153	212	20,941	1.00%
2	21,155	103	21,052	0.49%
1	21,154	37	21,117	0.18%
합계	**211,530**	**4472**	**207,058**	**2.11%**

Accuracy Ratio (AR) 기반 등급체계 평가

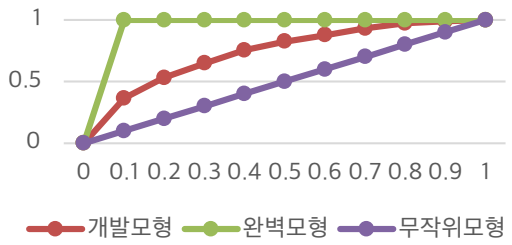


PULSE SCORING MODEL (PSM) (3/3)

모집단 전체가 아닌 무작위 추출을 통한 1, 2, 3차에 걸친 실험에도 유사하 수준의 결과가 지속적으로 도출되어 점수 체계 안정성이 강화되고 있습니다.

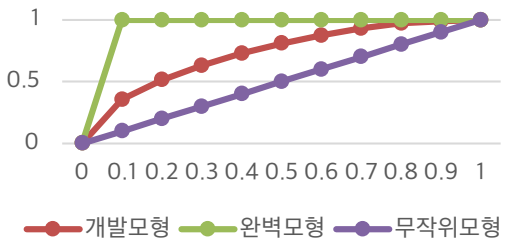
1차 시도 (AR = 53.2%)

분위	표본	부도 여부		부도율
		Y	N	
10	4,936	381	4,555	8.36%
9	4,962	174	4,788	3.63%
8	4,980	124	4,856	2.55%
7	5,058	112	4,946	2.26%
6	5,041	72	4,969	1.45%
5	5,038	55	4,983	1.10%
4	5,051	58	4,993	1.16%
3	5,058	41	5,017	0.82%
2	4,889	19	4,870	0.39%
1	4,987	11	4,976	0.22%
합계	**50,000**	**1,047**	**48,953**	**2.14%**



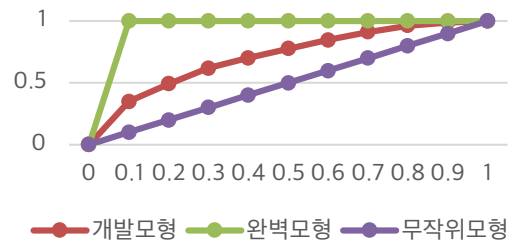
2차 시도 (AR = 51.2%)

분위	표본	부도 여부		부도율
		Y	N	
10	4,982	375	4,607	8.14%
9	5,045	166	4,879	3.40%
8	5,036	122	4,914	2.48%
7	5,016	105	4,911	2.14%
6	5,013	84	4,929	1.70%
5	4,993	68	4,925	1.38%
4	4,942	60	4,882	1.23%
3	4,946	43	4,903	0.88%
2	4,905	18	4,887	0.37%
1	5,122	12	5,110	0.23%
합계	**50,000**	**1,053**	**48,947**	**2.15%**



3차 시도 (AR = 47.8%)

분위	표본	부도 여부		부도율
		Y	N	
10	4980	364	4,616	7.89%
9	5043	148	4,895	3.02%
8	4871	133	4,738	2.81%
7	5003	85	4,918	1.73%
6	4915	78	4,837	1.61%
5	5095	71	5,024	1.41%
4	4976	69	4,907	1.41%
3	5066	53	5,013	1.06%
2	5061	27	5,034	0.54%
1	4990	12	4,978	0.24%
합계	**50,000**	**1,040**	**48,960**	**2.12%**



향후 개선 포인트

본 결과를 바탕으로 공공 데이터를 활용한 신용평가 가능성을 확인하였으며, 향후 데이터 항목 확대와 기업군 식별력 강화, 그리고 기준 상세화를 통한 모형 고도화가 충분히 가능할 것으로 판단합니다.

대안 데이터의 발생 여부 및 최신성이 법인 부도 예측에 유의미한 변수라는 것을 확인하였으며 고도화 가능

"강화"	"강화"	"강화"
데이터 유형	**기업군 식별**	**기준 상세화**
• 현재 당사가 보유한 전체 대안 데이터 약 300개 중 1/3 가량만 활용한 상태이며 이후 추가 편입 • 향후, 추가 데이터 투입과 파생 데이터 생성을 통해 보다 다양한 영역에서 모형 정교화 • **변수 항목을 대규모로 추가됨에 따라 분석 가능 기업 표본이 큰 폭으로 확대되고 예측력 필연적으로 증가**	• 현재 IBK 중소기업은행 제공 차주 표본 중 약 10% 수준이 미식별되어 본 실험에서 제외 • 누락된 10% 기업군의 부도율이 3% 이상으로 이를 분석할 경우 유의미한 Insight 도출 예상 • **당사 Hubble Database 2.0 Alpha 패치가 7월에 적용 예정인 바, 이에 대한 개선 가능**	• 현재 부도 판단은 IBK 중소기업은행 측이 제공한 약 25개 세부 기준에 근거하며 세부 유형 식별 불가 • 부도 사유의 경중과 성격이 상이하다고 판단하며, 이는 예측 모형의 Noise 작용 가능 • **따라서, 향후 부도 세부 유형에 따른 맞춤화 분석 시, 예측 정교성을 향상할 수 있다고 판단**

Agenda

기술 검증 추진 배경

공공 데이터 유의성 검증

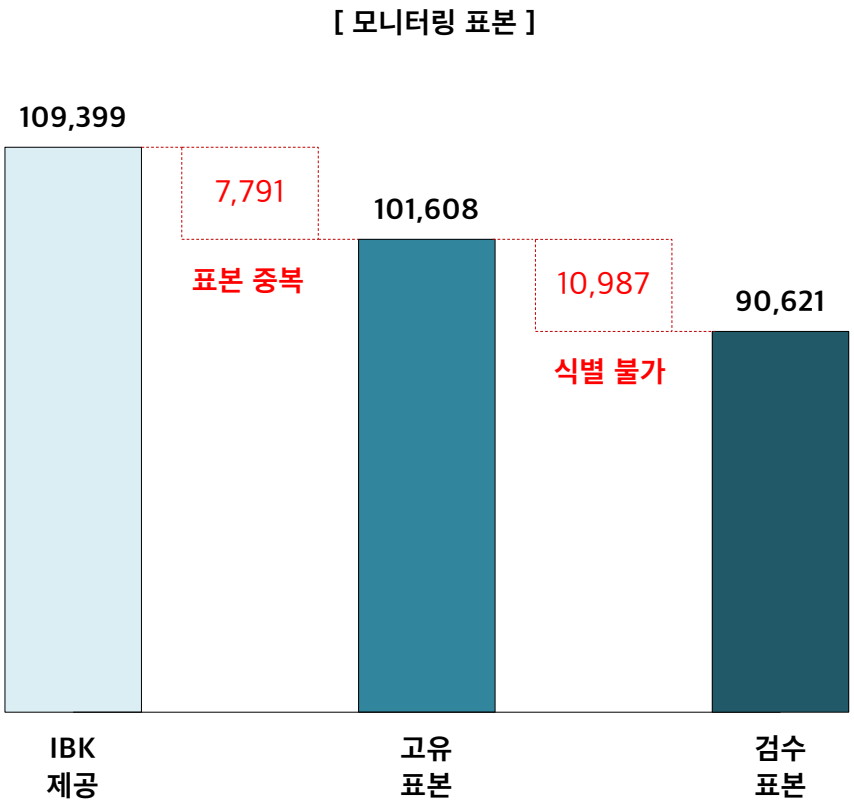
신용평가 (예측) 모형 검증

현장 테스트 결과

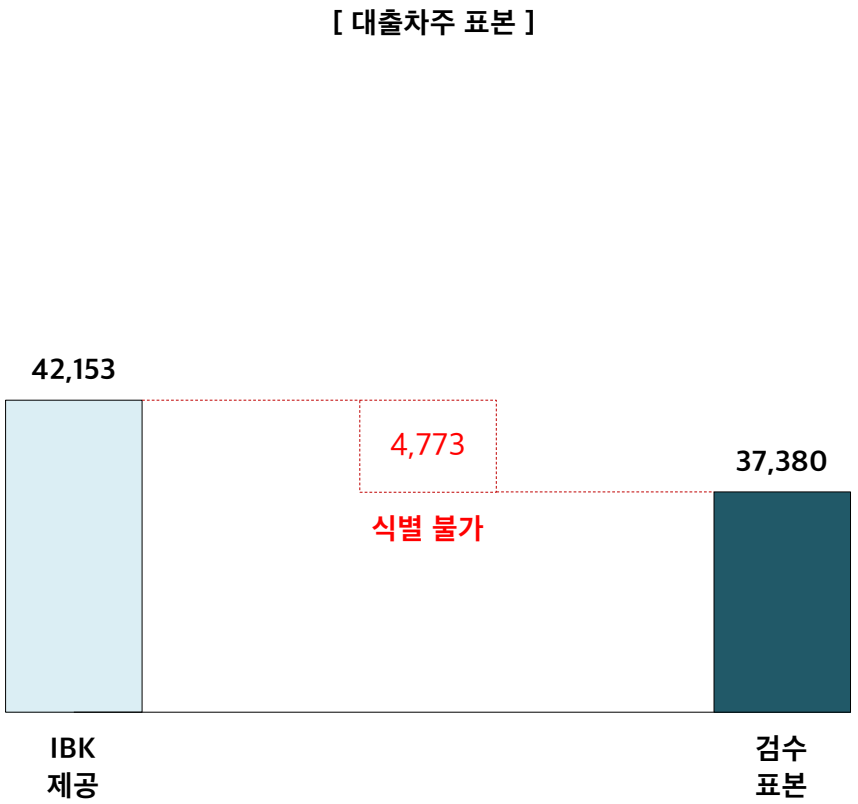
검증 대상 표본

최종 검증을 수행하기 위하여 IBK 중소기업은행으로부터 현장에서 테스트 표본을 전달 받았고, 검수 대상은 모니터링 표본 측면에서 90,621개 / 기업 차주 표본 측면에서 37,380개로 확정되었습니다.

IBK 1st Lab 최종 검수 **표본** 선정 결과



IBK 1st Lab 최종 검수 **기업** 선정 결과



검증 예측 방식

공공 데이터 기반의 예측 성능 평가를 위하여 이전 단계에서 선정한 표본 대상으로 Radom Forest 기반의 예측을 진행하며 혼돈행렬 (Confusion Matrix) 기반의 지표를 도출합니다.

AI 부도 예측 모형 개발 프로세스

변수 선정

- 내부 기술 검증 단계에서 데이터 유의성 검사를 통과한 30개 지표를 독립 변수로 유지하여 인공지능 모형에 투입

알고리즘 채택

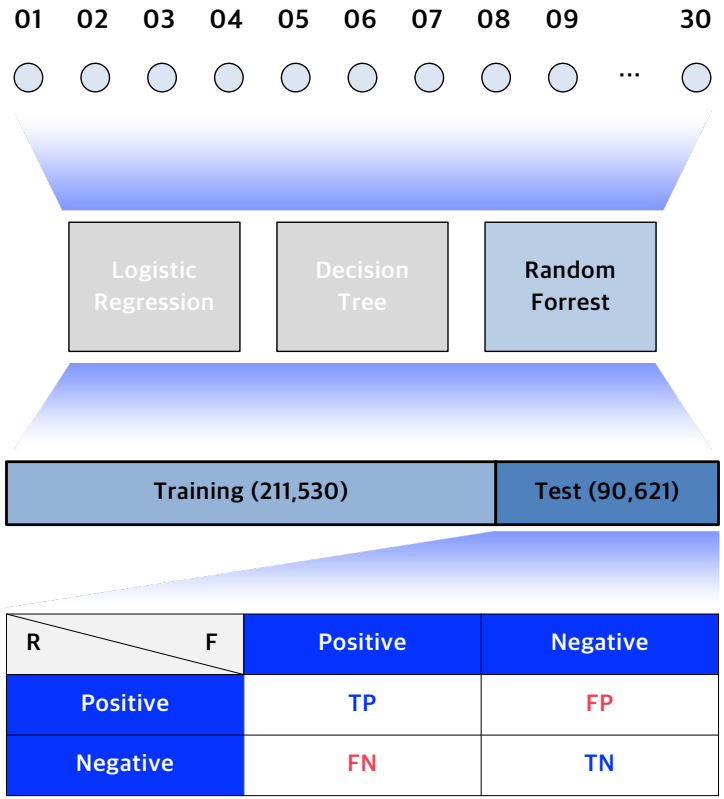
- Logistic Regression, Decision Tree 두 가지 모형의 경우 예측력이 제한적인 바, Random Forest 중심의 분석 수행

표본 구성

- 이전 1차 검수 작업에서 확보한 211,530개 표본을 대상으로 학습하고, 금번 최종 검수에서 이관 받은 90,621개 테스트 표본에 적용

성능 평가

- Confusion Matrix (혼돈행렬) 방식을 통해 성능 평가 진행하며, 이를 통해 Accuracy, Precision, Recall, F1, ROC_AUC 지표 도출



R \ F	Positive	Negative
Positive	TP	FP
Negative	FN	TN

부도 예측력 평가 결과 (1/2)

주요 인공지능 모형에 따른 분석을 완료하였으며 혼돈행렬 (Confusion Matrix) 상의 결과는 아래와 같고, 이중 Random Forest 방식이 가능 높은 성능을 기록하였습니다.

일차 검수 Random Forest

		예측		
		정상	부도	총계
실제	정상	51,582	222	51,804
	부도	402	677	1,079
	총계	51,984	899	52,883

최종 검증 Random Forest

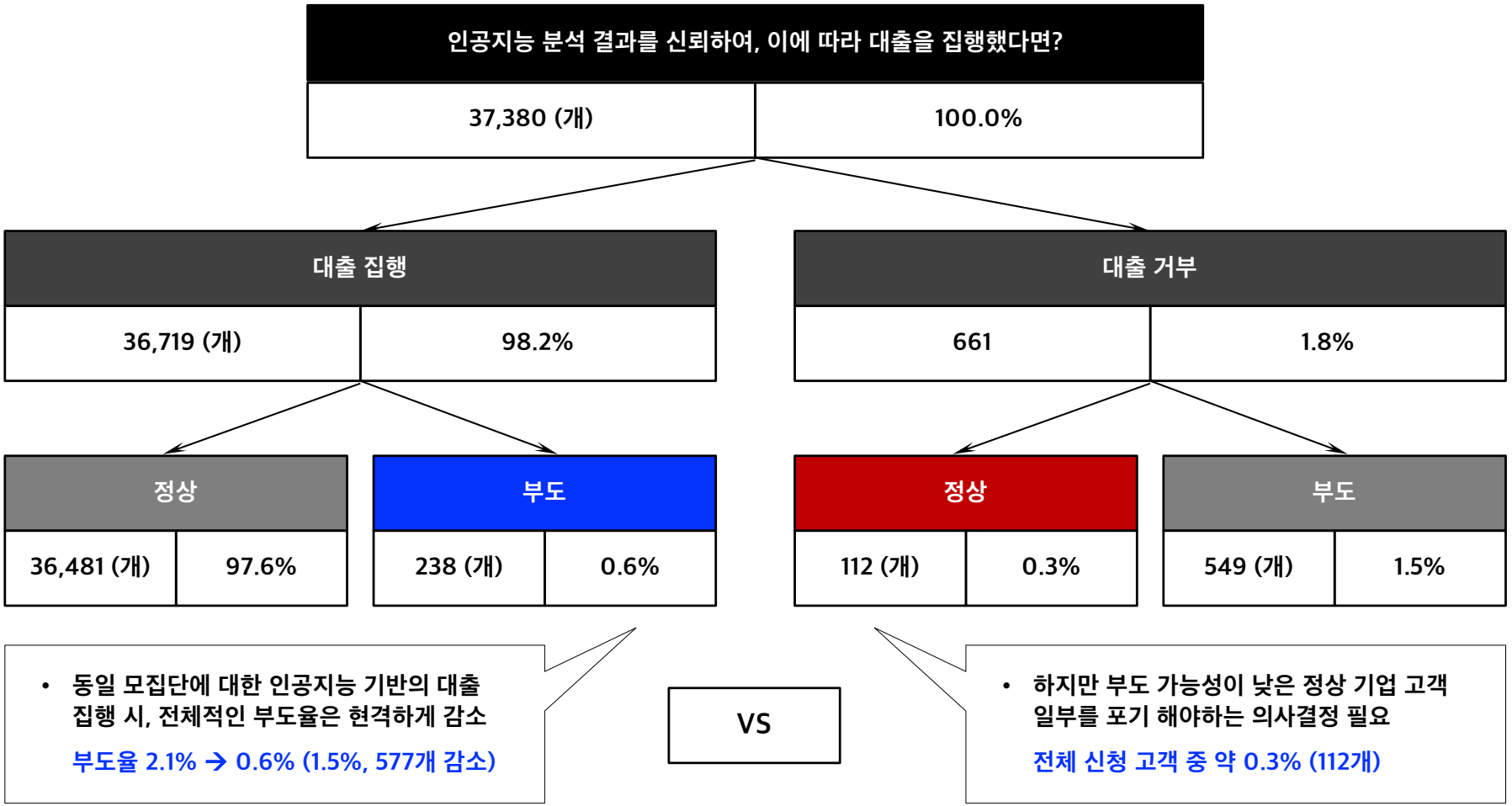
		예측		
		정상	부도	총계
실제	정상	88,441	272	88,713
	부도	577	1,331	1,908
	총계	89,018	1,603	90,621

- 금번 최종 학습의 경우, 이전 검증 단계에서 이관 받은 20만 건 이상을 학습할 수 있어서, 보다 정교한 예측이 가능
- 또한, 자체 분석 결과 이전 Sample Set 및 금번 검증 Set 간 중복 표본이 약 78% 존재하여 이에 대한 이중 학습

모형 성능 지표	점수		모형 성능 지표	점수
Accuracy	98.80%	+ 0.26%	Accuracy	99.06%
Precision	75.31%	+ 7.72%	Precision	83.03%
Recall	62.74%	+ 7.02%	Recall	69.76%
F1 Score	68.45%	+ 7.37%	F1 Score	75.82%
ROC_AUC	81.16%	+ 3.57%	ROC_AUC	84.73%

부도 예측력 평가 결과 (2/2)

실제 해당 모형에 따른 대출 의사결정을 실행할 경우, 기존 표본에서 부도 차주를 577개를 감축할 수 있으며 부도율이 2.1% 수준에서 1.5% 하락한 0.6% 수준으로 유지할 수 있습니다.



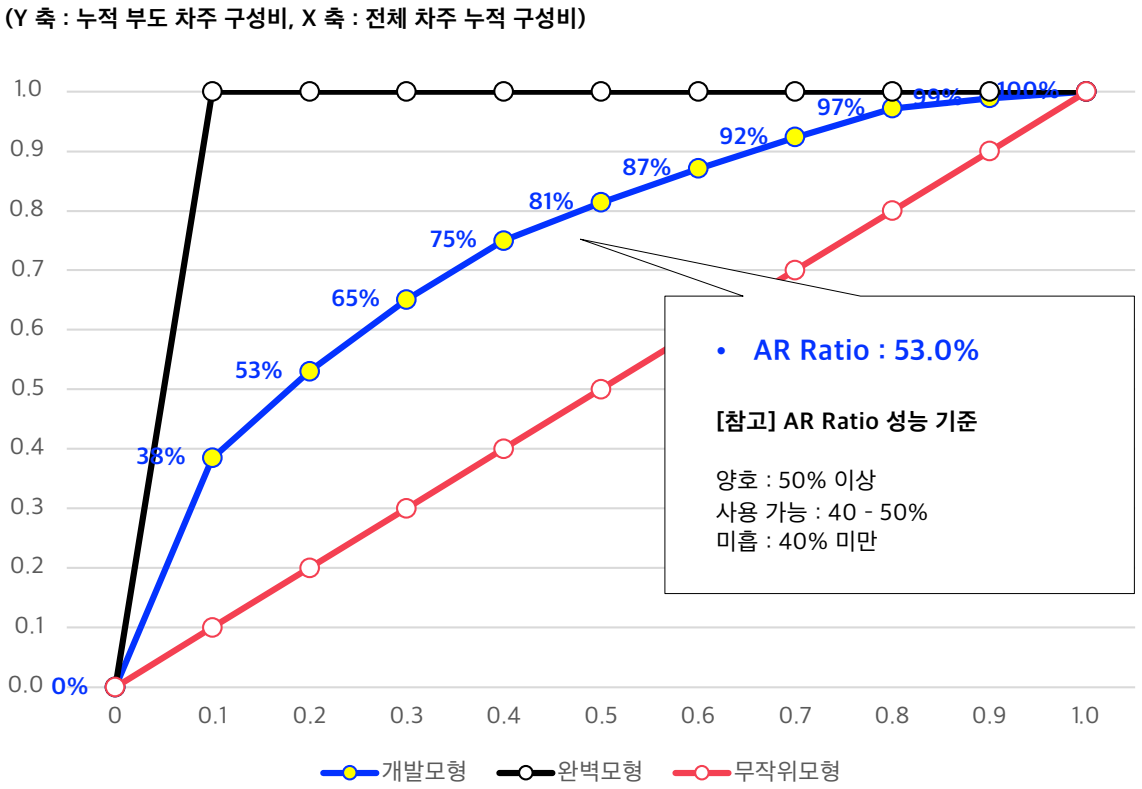
PULSE SCORING MODEL

최종 Test 표본에 근거한 PULSE 기반의 등급 분포도는 이하와 같으며, 최종적인 AR 점수는 53% 수준으로 도출되었습니다.

신용등급 분포표

분위	표본	부도 여부		부도율
		Y	N	
10	8,985	734	8,251	8.17%
9	9,014	278	8,736	3.08%
8	9,133	230	8,903	2.52%
7	9,092	189	8,903	2.08%
6	9,021	122	8,899	1.35%
5	9,077	108	8,969	1.19%
4	9,158	101	9,057	1.10%
3	9,145	92	9,053	1.01%
2	9,175	33	9,142	0.36%
1	8,821	21	8,800	0.24%
합계	**90,621**	**1,908**	**88,713**	**2.11%**

Accuracy Ratio (AR) 기반 등급체계 평가



DV