

## 1 추진배경

- 초지능·초연결·초융합 시대에 인공지능(Artificial Intelligence; AI)은 빅데이터 처리와 혁신적 서비스 개발의 필수·핵심 기술입니다.
  - 특히, 고품질 데이터가 풍부한 금융분야는 빅데이터·인공지능 활용의 잠재력이 매우 큰 분야입니다.
- 이에 금융위원회는 지난 '22.8월 금융권의 AI 활용을 지원하기 위하여 「금융분야 인공지능 활용 활성화 및 신뢰 확보 방안」을 발표하고, 세부과제를 착실히 추진하고 있습니다.

### ※ 「금융분야 인공지능 활용 활성화 및 신뢰 확보 방안('22.8월)」 주요 내용

#### ▶ 양질의 빅데이터 확보 지원

- ① 금융 AI 데이터 라이브러리 구축, ② 금융권 협업을 통한 데이터 공동 확보, ③ 데이터전문기관 추가 지정

#### ▶ AI 활성화를 위한 제도 정립

- ④ 금융 AI 개발·활용 안내서 발간, ⑤ 설명가능한 AI 요건 마련, ⑥ 망분리 및 클라우드 규제 개선

#### ▶ 신뢰받는 AI 활용 환경 구축

- ⑦ 금융 AI 테스트베드 구축, ⑧ AI 기반 신용평가모형 검증체계 마련, ⑨ AI 보안성 검증체계 구축, ⑩ AI를 활용한 효율적 감독체계 구축

- 동 방안에 따라 신뢰받는 AI 활용 환경을 구축하기 위해 관계기관과의 지속적인 협의 등을 통해 「AI 기반 신용평가모형 검증체계」와 「금융분야 AI 보안 가이드라인」을 마련하였습니다.

## 2 AI 기반 신용평가모형 검증체계 및 AI 보안 가이드라인 주요내용

### ① AI 기반 신용평가모형 검증체계 마련

- 신용평가는 금융소비자의 금융거래 조건 등에 영향을 미치며 금융회사의 리스크 관리에도 밀접한 관련이 있어 신용평가의 투명성과 공정성을 확보할 필요가 있습니다.
  - 이에 CB사의 기존 전통적 신용평가모형에 대해서는 신용정보법에 따라 '20년부터 「개인신용평가체계 검증위원회\*(신정원)」에서 검증하고 있습니다.
    - \* (구성) 위원장(신용정보원장), 학계·연구계, 금융계, 법조계, 소비자보호 전문가 6인
- 최근 신용정보회사(Credit Bureau; CB) 등은 AI 방법론을 이용한 신용평가모형을 개발·활용하고 있습니다.
  - AI 신용평가모형은 전통적인 신용평가모형에 비하여 더 다양하고 많은 데이터\*를 평가항목으로 반영할 수 있어 비금융·비정형 데이터\*\* 활용이 용이하며, 예측력·변별력이 우수하다는 장점이 있으나,
    - \* 전통적인 신용평가모형은 연체이력, 금융거래실적, 금융부채규모 등 금융정보 위주의 제한된 데이터만을 평가에 반영
    - \*\* (예시) 통신료, 건강보험·국민연금 보험료 납부이력, 온라인 쇼핑 구매 패턴, 모바일 앱 활용 패턴, 개인사업자의 점포가 속한 상권정보 등
  - 복잡한 알고리즘을 사용하기 때문에 평가결과에 대한 직관적인 해석과 설명이 다소 어렵다는 한계 등이 있으므로
  - AI 신용평가모형의 경우 신뢰성 확보와 소비자 보호를 위해 전통적인 신용평가와 같이 객관적인 검증체계를 마련할 필요성이 제기되었습니다.
- 이러한 측면에서 연구용역, 전문가 TF, 「개인신용평가체계 검증위원회」의 심도있는 논의를 거쳐 CB사에 대한 「AI 기반 신용평가모형 검증체계」를 마련하였습니다.

□ 「AI 기반 신용평가모형 검증체계」가 기존 전통적인 개인신용평가모형에 대한 검증체계와 차별화되는 주요내용은 다음과 같습니다.

① (데이터 관리) 신용평가에 활용하는 비금융·비정형 데이터에 대한 적절한 관리체계를 구축하였는지 점검합니다.

- 비금융·비정형 데이터를 신뢰성 높은 출처로부터 수집하고 데이터의 일관성·정확성\* 등을 주기적으로 확인합니다.

\* (예시) 개인사업자 매출 데이터가 전월 대비 10% 이상 변동시 동 매출 데이터 오류가능성 재점검

② (모형 선택) 다양한 AI 알고리즘의 특징과 장단점을 고려하여 신용평가에 최적화된 모형을 선택했는지 모형 선정 과정을 점검합니다.

- CB사는 다양한 알고리즘\* 중에서 하나를 선택하거나 다양한 알고리즘 결과를 조합하여 모형을 개발하는 만큼, 알고리즘 선정 목적, 변수 선정 과정 등 모형 개발의 상세 과정을 확인합니다.

\* AI 신용평가모형의 알고리즘은 크게 '신경망', '의사결정나무' 계열 등으로 구분되며, 다양한 세부 알고리즘이 존재

③ (설명 가능성) 금융소비자에게 신용평가모형 및 신용평가 결과 등을 충분히 설명할 수 있는지 확인합니다.

- CB사는 평가결과에 대한 설명의무\*가 있는 만큼, 설명가능한 AI 기법의 적용 여부, 해당 기법을 통한 모형의 해석 가능성 등을 점검합니다.

\* 개인은 신용평가 결과, 평가 기준, 평가에 이용되는 정보의 내용 등에 대하여 설명을 요구 가능(신용정보법 제36조의2)

④ (모형 성능) AI 방법론의 특성을 반영하여 모형의 변별력·안정성 등 통계적 유의성을 점검합니다.

- 과적합(overfitting\*) 가능성 점검, 학습·검증·테스트 데이터의 유사성\*\* 확인 등 AI 모형에 특화된 성능 확인 방법을 마련했습니다.

\* 모형이 학습 데이터에 과도하게 의존하는 것을 의미 → 학습 데이터에 대해서는 정확한 예측을 제시하지만 미학습 데이터에 대해서는 부정확하게 예측할 우려

\*\* 학습 데이터와 검증·테스트 데이터의 특성이 지나치게 다른 경우, AI 모형의 성능이 낮은 원인이 AI 모형의 문제인지 데이터 특성이 달라서인지 파악하기 곤란하여 모형 검증의 신뢰성이 낮아짐

## ② 금융분야 AI 보안 가이드라인 마련 [별첨]

□ AI 서비스 활용 확대와 더불어 개인정보 유출, 학습 데이터 조작 등 다양한 보안위협이 발생할 우려가 증가하고 있습니다.

○ 보안성을 충분히 확보하지 못하는 경우, AI서비스가 오작동\*하거나 악의적인 공격에 노출될 가능성이 있습니다.

\* (예시) 정신의학과 챗봇이 정식 출시 전 모의 환자에게 자살을 권유한 사례

□ 특히, AI 개발과정에서 현장 실무자가 유의해야 할 구체적인 보안위협과 대응방안을 다루는 세부적인 안내서가 필요하다는 지적이 제기되었습니다.

□ 이에 금융 AI 서비스 개발 실무자가 활용할 수 있는 개발단계별 세부 보안 안내서인 「금융분야 AI 보안 가이드라인」을 마련하였습니다.

○ 동 가이드라인은 ①AI 모델 개발단계별 보안 고려사항과 ②AI 챗봇 서비스에 대한 보안성 체크리스트를 제시하고 있습니다.

□ AI 모델 개발단계별 주요 보안 고려사항은 다음과 같습니다.

① (학습 데이터 수집) 오염된 데이터\*를 학습하여 발생할 수 있는 보안 문제와 성능 저하 등을 방지하고, 데이터 관련 공격·장애 발생시 그 원인을 파악할 필요가 있습니다.

\* 악의적 목적으로 변조한 데이터 등

→ 신뢰성 높은 출처로부터 학습 데이터를 수집하고, 데이터 출처 및 수집 시점 등을 파악할 수 있는 데이터 관리체계를 구축해야 합니다.

② (학습 데이터 前처리) 수집한 데이터를 학습에 적합한 형태로 가공하여 AI 모델의 품질과 보안성을 높일 필요가 있습니다.

→ 이상치(outlier)를 확인·처리하고, 적대적 예제\* 생성·학습 등을 통하여 AI 모델에 대한 적대적 공격\*\*을 예방하여야 합니다.

\* AI 모형이 잘못된 예측을 하도록 의도적으로 변조한 데이터

\*\* 적대적 예제를 활용하여 AI 모형이 잘못 판단하도록 조작하는 공격

③ (AI 모델 설계·학습) 잠재적 공격자가 AI 모델에 대한 적대적 공격 등을 쉽게 수행할 수 없도록 AI 모델을 구성할 필요가 있습니다.

→ 잠재적 공격자가 AI 모델에 대한 정보를 쉽게 유추할 수 없도록 지나치게 단순한 설계를 지양하고, AI 모델을 세부 변형하는 보안 기법\* 등을 적용하여야 합니다.

\* (예시) AI 모형의 성능은 유지하면서 AI 모형의 구조를 변경

④ (AI 모델 검증·평가) 학습을 완료한 AI 모델이 잠재적 공격 또는 개인정보 유출 등으로부터 안전한지 보안성을 검증할 필요가 있습니다.

→ AI 모델을 대상으로 선제적인 적대적 공격을 수행하여 AI 모델이 공격을 탐지·방어할 수 있는지 확인하여야 합니다.

→ AI 모델의 입·출력\* 횟수를 제한하여 잠재적 공격자가 AI 모델에 대한 정보를 수집하기 어렵게\*\* 하여야 합니다.

\* 입력 : 소비자가 AI 모델에 제공하는 문자나 숫자 등(예: 챗봇에 입력하는 질문)

출력 : AI 모델이 산출하는 값(예: 챗봇의 대답)

\*\* 공격자는 AI 모델의 입·출력값을 바탕으로 모델의 정보를 유추하고 공격을 시도

→ 공격자가 수집할 수 있는 값 제한 필요(예: 챗봇 대답을 1분에 10회 미만으로 제한)

→ AI 모델을 통하여 개인정보가 출력되는 경우, 개인정보가 타인에게 노출되지 않도록 하여야 합니다.

□ 금융분야에서 AI가 가장 활발히 사용되는 분야 중 하나인 챗봇서비스에 대해서는 「AI 챗봇서비스 보안성 체크리스트」를 별도로 마련하였습니다.

○ AI 챗봇서비스의 보안성 확보에 필요한 사항들을 체크리스트 형태로 구체화해 실무자가 챗봇서비스 개발시 쉽게 활용할 수 있을 것으로 기대됩니다.

※ (참고) 체크리스트 문항 예시

분류	소분류	연번	보안성 체크리스트
공통	입력제한	6	개인정보를 처리하지 않는 챗봇의 경우 입력창에 개인정보를 입력하지 않도록 이용자에게 사전에 안내하는가?