금융AI의 활용과 금융소비자 보호: 차별금지, 설명요구권, 6대 판매규제 준수를 중심으로

박상철*

I. 논의의 배경

금융이 향후 AI 활용이 가장 활발한 분야 중 하나가 될 것이라는 전망은 충분한 근거가 있다. 현시점에서의 대표적인 AI의 구현방식인 기계학습(machine learning)은 회귀분석(regression analysis) 등 전통적 통계기법들이 고도화된 것이다. AI의 산업적 활용이 본격화되기 전부터, 금융산업에서는 과거의 데이터로 미래를 예측하는 계량적 기법이 널리 활용되었다. 은행업・신용카드업・신용정보업에서의 신용평가(credit scoring), 금융투자업에서의 퀀트(quant) 기법, 알고트레이딩(algo trading) 및 전자적 투자조언장치(robo-advisor), 보험업에서의 계리적(actuarial) 기법과 계약심사(underwriting), 신용카드업・보험업・은행업을 중심으로 한 이상거래탐지(fraud detection)가 대표적 활용 분야이다. 다만 전통적인 계량모델들은 대부분 선형모델(linear model)로서 비선형적 분포를 보이는 데이터에 대한 과소적합(underfitting)의 문제가 있어, 기계학습의 발전 이후 금융기관들은 커널화된 서포트벡터머신(kernalized SVM), 결정나무(decision tree), 심층학습(deep learning) 등 다양한 비선형모델을 통해 몸매에 맞는 옷을 재단하듯 편이(bias)를 줄이며 데이터의 다양한 패턴을 보다 적확히 추론하고 있다. 역으로 전통적 계량모델들은 과거 데이터에 대한 과적합(overfitting) 때문에 도리어 예측력이 떨어지고 이례적 샘플(outlier 또는 "black swan")에 취약하였으므로, 데이터셋의 훈련셋(train set)과 테스트셋(test set)으로의 분리, 교차검증(cross-validation), 정규화(regularization), 배깅(bootstrap aggregation)을 포함한 다양한 기법을 통해 데이터의 분포에 내재한 분산(variance)을 통제하고 예측정확도를 향상시키고 있다.[1] 그러나 기계학습 기법은 전통적 계량모델과 본질적 차이가 있

* 서울대 법학전문대학원 조교수. 필자는 서울대학교 산학협력단의 2021년 금융분야 인공지능 활성화를 위한 가이드라인 등 마련을 위한 금융위원회 연구용역 수행 과정에서 금융에서의 대화형 에이전트 등의 활용에 대하여 연구책임자이신 고학수 교수님으로부터, 해외 정책 동향에 대하여 공동연구자이신 김병필 교수님으로부터 많은 배경지식을 얻었고 이 연구에 큰 도움이 되었음을 밝힙니다.

1) 예컨대 미국 금융기관들은 대출을 신청한 기업들의 도산 확률을 예측하기 위하여 Altman (1968)이 실제 데이터로 선형판별분석(linear discriminant analysis; LDA) 모델을 훈련시켜 고안한 Altman Z-score (Edward I. Altman, *Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy*, 23(4) J Finance 189 (1968))를 널리 쓰고 있었는데, Altman 스스로가 2017년 저자로 참여한 논문에서 랜덤포레스트(random forest) 모델 사용 시 선형모델에 입각한 기존의 Z-score에 비해 예측정확도를 뚜렷이 향상시킬 수 있음을 밝히기도 하였다(Flavio Barboza, Herbert Kimura & Edward Altman, *Machine Learning Models and Bankruptcy Prediction*, 83 Expert Syst Appl 405 (2017)).

다기보다는 연속선상에 있기 때문에, 이를 기존에 많이 활용했던 금융산업에서 기계학습 또한 그대로 많이 활용될 것으로 예상할 수 있다. 규범적 접근에 있어 양자 간 차이를 실제보다 과대평가해서도 아니 된다.

다만 최근의 AI의 발전 및 사업화 양상을 보면 전통 모델로부터의 질적 차이를 수반하는 혁신도 분명 존재한다. 예컨대 데이터 가용성(data availability)의 비약적 향상에 힘입어 온라인플랫폼 입점 사업자 등의 다양한 비금융데이터를 바탕으로 대출심사를 하는 대안신용평가시스템(ACSS)이 도입되었고,[2] 본인신용정보관리업(마이데이터)의 활성화는 특정 개인을 기준으로 프로파일링된 양질의 데이터셋의 가용성을 향상시킬 수 있다. 나아가 자연어(natural language)나 영상 등 다량의 비정형데이터(unstructured data)를 순환신경망(RNN)이나 합성곱신경망(ConvNet) 등 향상된 알고리듬을 바탕으로 처리하여 사람의 인지(cognition)를 모사할 수 있게 됨으로써 예측 기법은 더 고도화되었다. 대표적인 사례는 미국의 렌도(Lenddo) 등 사회관계망서비스(social media) 기반 신용평가 및 소액대출 업체들이고, 이들을 모델 삼아 개정 신용정보법(2020. 8. 4. 시행)은 신용정보회사 등이 "신용정보주체가 스스로 사회관계망서비스 등에 직접 또는 제3자를 통하여 공개한 정보"의 경우 신용정보주체의 동의 없이 수집할 수 있도록 하여(제15조 제2항 제2호 다목) 소셜미디어 텍스트 마이닝에 기한 신용평가를 활성화하였다. 특히 P2P 금융과 관련하여 챗봇, 소셜미디어, 심리테스트 등 다양한 데이터들이 신용평가에 활용되고 있다.[3] 텍스트 마이닝을 통해 트렌드를 파악하여 주식・외환 시세 예측, 고객관리 등에 활용하는 기법도 꾸준히 시도되고 있다.[4]

다른 한편, 2007년 서브프라임 모기지 사태 이후의 대침체(Great Recession), 2008년 키코(KIKO) 사태, 2012년 저축은행 후순위채 사태, 2019년 파생결합증권・펀드(DLS/DLF) 사태, 2020년 옵티머스・라임 사태 등을 거치며 (바젤III 등에 기한) 건전성 규제(prudential regulation)와 (6대 판매규제 등) 소비자보호 규제(codes of conduct regulation)가 숨가쁘게 확장되었고, 이로 인해 규제준수(compliance)의 효율화를 위한 AI의 활용(CompTech)에 대한 수요가 기존 계량적 기법의 고도화를 위한 AI의 활용 수요에 못지않게 절실해진 측면이 있다. 특히 소비자보호 규제준수는 특성상 수량화・구조화된 데이터에 기한 모델로 구현하기 어려워 고객 상담 녹취 파일의 텍스트변환 및 분석, 챗봇(chatbot) 등 대화형 에이전트(conversational agent)의 활용 등 다양한 자연어 처리 기법들이 시도되고 있다. 2021. 3. 25.부터 최초 시행된 금융소비자보호법상 6대 판매규제가 이를 가속화하고 있음은 물론이다.

결국 금융AI에 있어 금융소비자 보호의 문제는 AI로부터의 금융소비자 보호(protection from AI)와 AI에 의한 금융소비자 보호(protection by AI)의 양 측면에서 접근해야 한다. AI로부터의 보호의 경우 금융 분야에서의 기계학습 기법의 활용과 관련하여 우려와 규제론이 제기되나, 기존에 활발히 활용되었던 계량적 기법들에 비해 어떠한 증분적 위험(incremental risk) 뿐 아니라 이를 상쇄하는 어떤 편익이 생기는지 여부를 차분히 파악하는 것에서 출발해야 한다. 이를 통해 금융소비자보호법, 신용정보법 등 관련 법령에 규정된 차별금지와 설명요구・이의제기권 조항이, AI모델의 공정성과 투명성과 관련된 보다 광범위한 논변들의 맥락에서, 금융AI에 구체적으로 어떤 방식으로 적용될 수 있을지의 문제를 살

2) 네이버파이낸셜과 미래에셋캐피탈이 2020. 12.부터 전자가 네이버 스마트스토어에 입점한 사업자들의 매출 흐름, 반품율, 단골 고객 비중, 고객 평점, 고객 문의 응대 속도 등 비금융데이터를 바탕으로 기계학습 기반 대출심사를 하면, 후자가 이를 바탕으로 대출을 해주는 상품을 출시한 사례 참조(매일경제, 2021. 1. 29., "네이버파이낸셜, 스마트스토어사업자 대출 확대").

3) 최수만/전동화/오경주, "P2P 플랫폼에서의 대출자 신용분석 사례연구: 8퍼센트, 렌딧, 어니스트 펀드," 지식경영연구 제21권 제3호(2020. 9.), 한국지식경영학회, 238-241면.

4) B. Shravan Kumar & Ravi Vadlamani, *A Survey of the Applications of Text Mining in Financial Domain*, 114 Knowl-Based Syst 128 (2016).

펴본다. AI에 의한 보호 문제의 경우 금융소비자보호법상 6대 판매규제가 정보 비대칭성 등 감소라는 편익을 위해 전체 소비자와 금융상품판매업자등의 거래비용 증가 등 사회적 비용을 유발할 우려가 있는 상황에서, AI의 활용을 통해 이러한 비용을 절감하고 의도한 소비자보호 또한 효과적으로 달성할 수 있도록 어떤 제도적 설계를 해야 하는지를 중점적으로 검토한다.

II. AI로부터의 금융소비자 보호: 차별금지와 설명요구권을 중심으로

1. 개요

2020. 12. 10.부터 시행된 지능정보화 기본법은 정부가 AI 개발자・공급자・이용자가 준수하여야 하는 윤리원칙(인간의 존엄과 가치의 존중, 공공성・책무성・통제성・투명성 등)을 정한 지능정보사회윤리준칙을 제정하여 보급할 수 있다고 규정하고(제62조 제1항, 제4항), 이에 따라 과학기술정보통신부가 2020. 12. 23. AI윤리기준을 마련하였다. 이는 공정성(fairness), 투명성(transparency), 책무성(accountability)을 골자로 한 AI윤리(AI ethics) 또는 신뢰가능한 AI(trustworthy AI)의 기준들이 전 세계적으로 마련되고 있는 추세를 따른 것이다. 이들의 논지는 AI모델이 불투명(opaque)하고 설명불가능(inexplicable)한 소위 블랙박스 모델(black-box model)[5]인 경우 분류 대상인 인간의 자율성을 침해할 수 있고,[6] 데이터에 내재한 편향(bias)으로 인해 차별적일 수 있으며,[7] 제3자 등에 위해를 가하더라도 책임소재의 추적이 어려울 수 있으므로,[8] 이에 대처하기 위한 규범 체계가 필요하다는 것이다. 이러한 논변에 대해서는 비판적 접근이 필요하나, 리스크 기반 접근(risk-based approach)을 통해 고위험 AI(high-risk AI)을 식별하여 어느 정도 통제할 필요가 있다는 견해는 설득력이 있을 수 있다.[9] 구체적으로 (1) 교통, 에너지, 의료 등에 사용되는 로봇 등 자율시스템의 경우 공중안전을 위협하는 사고 리스크의 통제가,[10] (2) 거래시스템에 접속된 자동매매 에이전트의 경우 시장안정성에 대한 위해(플래시크래시(flash crash) 등 시장변동성의 증가, 쏠림현상, 시스템오류, 주문실수 등[11])의 통제가, (3) 중대한 혜택 또는 불이익의 대상자를 분류하는 모델(채용・입학전형・신용평가・사법 AI 등)의 경우 비차별성의 확보가 필요하고 그 전제로 어느 정도의 추적성(traceability) 내지 설명가능성(explainability)이 필요하다는 것이다.[12] 금융 AI와 관련해서는 두 번째 문제(자동매매

5) 국내외에서 "블랙박스 알고리듬"이라 표현하는 경우가 많으나 잘못된 용어이다. 알고리듬은 개발자가 완벽하게 이해할 수 있다. 개발자의 해석가능성(interpretability)이 떨어질 수 있는 것은 알고리듬이 아닌 알고리듬으로 훈련된 모델(model)이다.

6) Executive Office of the President, *Preparing for the Future of Artificial Intelligence*, 32 (2016). European Commission, *White Paper on Artificial Intelligence: a European Approach to Excellence and Trust*, COM(2020) 65 final, 12 (2020). High-Level Expert Group on AI, European Commission, *Ethics Guidelines for Trustworthy AI*, 13 (2019).

7) EOP, *supra* note 6 at 30-32. EC, supra note 6 at 11-12. AI HLEG, *supra* note 6 at 18-19.

8) EOP, *supra* note 6 at 30-32. EC, *supra* note 6 at 11. AI HLEG, *supra* note 6 at 19-20.

9) EC, *supra* note 6 at 17.

10) EC, *supra* note 6 at 17.

11) 한국거래소 시장감시위원회, 「알고리듬거래 위험관리 가이드라인」(2014) 참조.

12) 유럽집행위원회(EC)가 유럽의회(EP)에 2021. 4. 21. 제출한 "인공지능에 관한 통일규범(인공지능법)의 제정 및 일부 연합제정법들의 개정을 위한 법안"(Proposal for a Regulation laying down harmonised rules on artificial intelligence (Artificial Intelligent Act))은 (1) 잠재의식 조작, 저연령・장애인 취약성 악용, 사회신용도 분류, 공공장소 실시간 원격감시 AI는 금지하고, (2) 인증 대상 제품의 안전성 요소, 생체인증, 필수기반시설, 입학전형, 채용, 공공부조 심사, 신용평가, 긴급출동, 형사절차, 출입국관리, 사법 관련 AI는 고위험 AI로서 적합성평가(conformity assessment)를 거치도록 하고 데이터 거버넌스, 기록보존, 투명성, 통제성, 정확・견고・보안성 등 규제를 가하며, (3) 사람과의 상호작용, 감정인식, 생체범주화, 딥페이크 AI는 투명성 의무를 부과한다.

에이전트)도 중요한 토픽이나 이 글의 주제 범위를 벗어나므로 생략하고 세 번째 문제(분류모델)에 집중한다.

우리 금융시장에서 AI 기반 분류모델(classification model)은 다양하게 활용되고 있으나 (1) 계약심사 및 조건 산정(여신심사[13], 신용평가[14], 보험계약심사(underwriting)[15] 등), (2) 사기 또는 이상거래탐지(신용카드거래 승인,[16] 보험금 지급심사(손해사정(claims adjusting) 포함)[17], 은행・증권 이상거래[18] 등), (3) 추천 알고리듬(맞춤형 마케팅, 맞춤형 광고, 맞춤형 설명의무 이행 등)이 대표적 활용례로 보인다. 이러한 AI 기반 분류모델에서 공정성과 투명성이 확보되어야 한다는 논변의 난점은, 공공영역이 아닌 민간영역(private sector)에서, 단지 활용된 기술의 (때때로 비본질적인) 차이만으로, 공공영역에 요구되는 정도의 공정성과 투명성이 강제되어야 할 근거가 명확하지 않다는 것이다. 이 논리적 비약(logical leap)은 사고실험(thought experiment) 수준의 "블랙박스에 의한 자율성 침해론"만으로 메우기 어렵다. 인권규제의 민간영역으로의 확장을 지향하는 사회자유주의(social liberalism)의 기획이 (종래와 같이 "포괄적 차별금지법"과 같은 투박한 경성법(hard law) 논리에 머무르지 않고) AI윤리, ESG 등 세련되게 주조된 개념과 그에 입각한 연성법(soft law)을 매개로 관철되고 있는 세계적 추세의 일환이나, 결국엔 경성법으로 전환되는 시점에서 규제비용에 대한 점검이 필요하다. 물론 금융의 경우 공공영역이 아님에도 불구하고 시장지배력 여부를 묻지 않고 고도의 공익성을 전제한 법적 규율을 이미 받아 왔고 오히려 심화 중이며, 이것이 우리만의 현상이 아닌 세계적 추세인 점은 고민을 다소 덜어준다.[19] 그러나 특히 인권규제의 금융에 대한 전면적 확장, 특히 금융에 대한 (전력, 교통, 통신 규제를 방불케 하는) 보편적 서비스(universal service) 내지 공익산업(utility)에 가까운 차별금지 규제에는 다음과 같은 난점이 있음 또한 인지되어야 한다.

첫째, 일정한 차별금지사유에 기한 차별을 금하는 규제(금융소비자보호법 제15조, 대부업법 제9조의8, 신용정보법 제22조의4 제2항 제1호, 제3호 등)의 경우 상대적으로 엄격하게 적용・집행될 것으로 전망된다. 이 경우 해당 차별금지사유가 신용위험, 보험위험률 등과 밀접한 상관관계가 있으면, 고위험 소비자를 위한 보편적 서비스 제공을 의무화하고, 시장구조에 따라 손실 중 일부는 각 금융기관이 준조세(quasi-tax)처럼 떠안게 하면서, 다른 일부는 저위험 소비자들이 상호보조(cross-subsidization)(신용도나 저위험에 상응하는 우대 박탈)[20]하도록 하는 것과

13) 하나은행이 2020년경부터 자영업자(소호) 여신심사 시 기존의 정량지표에 기한 신용평가시스템(CSS)에 더하여 업종, 지역, 현황 등 비금융데이터에 기한 기계학습 모형을 결합하여 자동신용평가를 한 사례 참조(더벨, 2020. 9. 1., "하나은행, 소호대출 '드라이브' AI 심사 적용").

14) 네이버파이낸셜과 미래에셋캐피털의 대안신용평가시스템(주 2) 등 참조.

15) 삼성화재가 2019. 9.부터 장기보험에 AI 계약 심사 시스템을 도입해 장기인보험의 경우 심사자의 별도 확인 없이 전산심사만으로 가입을 승인하고, 장기재물보험의 경우 AI가 이미지 및 자연어처리를 통해 업종과 관리상태를 판단해 업무처리 시간을 단축한 사례 참조(연합뉴스, 2019. 10. 8., "삼성화재, 장기보험 계약 심사에 AI 도입").

16) 신한카드가 2017. 6.부터 해외 오프라인 결제에, 2018. 10.부터 국내외 온오프라인 결제 및 현금융통에 기계학습 기반 FDS 도입한 사례 참조(신한카드, 2018. 10. 10., "보도자료: 머신러닝 부정사용방지시스템(FDS) 구축 완료").

17) 한화생명이 2019. 12.부터 ConvNet에 기해 실손보험과 정액보험 보험금 청구 시 각 병원 허가 병상 수, 수술 가능 환자 수 등을 분석해 보험사기를 탐지하는 자동심사시스템을 도입한 사례 참조(매일경제, 2020. 12. 24., "인공지능 보험금 자동심사 도입... 비용 절감하고 고객 편의성 높여").

18) 신한은행이 2016. 8.경 핀테크 업체와 연계하여 심층학습 기반 FDS를 도입한 사례 참조(금융보안원, "국내외 금융권 머신러닝 도입 현황," 보안연구부-2017-024, 4면).

19) 예컨대 헌법재판소는 금융기관 임직원의 직무관련 수재 등 행위를 공무원의 수뢰죄와 같은 수준으로 가중처벌하도록 한 특정경제범죄법 조항에 대해 "금융기관은 사기업이지만, 국가경제와 국민생활에 중대한 영향을 미치는 업무를 담당하고 있고, 시장경제질서의 원활한 운용을 위해서는 투명하고 공정하게 그 기능을 수행"하여야 한다는 점을 근거로 합헌결정을 내렸다(헌법재판소 2012. 12. 27. 2011헌바217 결정, 헌법재판소 2017. 12. 28. 2017헌바193 결정).

경제적으로 동일해진다. 이는 정부가 금융소외계층을 지급보증, 또는 보험료나 자기부담금 지원 형태로 직접 지원하거나 그렇지 않더라도 (통신산업에서처럼) 원가구조가 우월한 선두사업자가 대표로 보편적 서비스를 제공하고 다른 사업자들로부터 현행원가 등을 기준으로 산정된 보편적역무손실보전금(universal service fund; "USF")을 걷어 손실을 보전받도록 하는 것(전기통신사업법 제4조)보다도 비효율적이고, 후천적 요인(특히 노력)으로 인해 위험이 낮은 소비자들에게 (차별의 기준 여하에 따라서는) 차별적이다. 근본적으로, 정부나 의회가 소득세제와 복지지출로써 해소해야 할 불평등을, 금융소비자들이 금융기관에 맡긴 예금이나 보험료 등을 마치 세금처럼 간주하며 자신의 지시대로 할당하도록 하는 방식으로 해결하려는 것은 문제가 있다. 더 심각한 것은 고위험 소비자들에게는 금리나 보험료율이 자신의 리스크에 비해 낮기 때문에 수요초과가 발생하여 인위적 신용배급(credit rationing)을 해야 하므로 정실주의(favoritism)가 횡행하거나, 역선택(adverse selection)으로 인해 금융시장 전체가 부실(lemon market)화될 수 있다는 점이다. 이는 건전성 규제가 지향하는 정책목표에 완전히 배치되는 결과이다.

둘째, 차별금지사유를 정해놓지 않고 부당하거나 정당한 사유 없는 가격차별(price discrimination)을 일반적으로 제한하는 경우(금융투자업법 제58조 제2항, 보험업법 제129조 제3호, 여신전문금융업법 제18조의3 제1항, 온라인투자연계금융업법 제11조 제4항 등)는 실무상 일물일가주의(Law of One Price)까지는 아니고 후술하듯이 합리적 혹은 정책적 이유로 차등을 상당히 허용하는 것으로 보이나, 차등이 제한되는 한에서 소비자들 간 상호보조를 막겠다는 명분하에 실제로는 프론트엔드(소매 단)에서의 가격경쟁과 소비자 전환을 저해하는 문제점이 있다.[21] 이는 금융재판매를 허용하고 기존 금융기관들의 백엔드(도매 단)에서 금융결제망 등 금융인프라를 세분화(unbundle)한 후 필요하면 재판매사업자에 대한 동등접속(equal access) 의무를 부과하여 프론트엔드에서의 시장진입 및 치열한 가격인하 경쟁을 유도하는 것보다 하수(下手)다. 시장압력을 거스른 소매가격 규제는 결국 유통망을 통한 리베이트 차별로 우회되는데, 이를 또 다른 땜질규제(regulatory patchwork)로 틀어막아야 하므로[22] 집행비용은 더 상승한다. 핀텍(fintech)의 발전은 애그리게이터(aggregator)를 중심으로 한 재판매사업의 허용 압력으로 작용할 것이고 본인신용정보관리업이 안착될 경우 그 흐름을 앞당길 수 있으며 소매가격 규제는 더 무색해질 것이다.

셋째, 차별금지조항이 "차별적 취급"을 넘어 "차별적 발화"의 금지, 즉 혐오표현(hatred speech)의 제재와 정치적 올바름(political correctness)을 위해 전용(轉用)되어야 한다는 목소리가 커질 것이다. 금융상품 판매 직원들 이상으로 6대 판매규제 등의 준수를 위해 도입한 챗봇 등 대화형 에이전트가 왜 사람 수준의 성인지감수성을 갖추지 못했냐는 문제제기가 늘 것으로 보인다. 차별적 발화가 차별적 취급에 선행되어 차별적 의도의 공공연한 증거(overt evidence)가 되는 경우를 제외하면, 발화 자체는 (공적 인물의 경우) 명예훼손이나 모욕으로, (비공적 인물의 경우) 개인정보나 프라이버시 침해로 다루어야 할 사안으로서 차별금지조항의 적용대상은 아님을 전제한다.

이러한 난점에도 불구하고 금융이 이미 공공부문에 준하여 규제받던 것을 주어진 것으로 받아들인다면,

---

20) Roger L. Pupp, *Community Rating and Cross Subsidies in Health Insurance*, 48(4) J Risk Ins 610 (1981).

21) U.K. Financial Conduct Authority, *Price Discrimination and Cross-Subsidy in Financial Services* (Occasional Paper No. 22), 32-33 (2016), https://www.fca.org.uk/publication/occasional-papers/op16-22.pdf.

22) 대표적인 예로는 보험업법상 보험요율 차별 금지(제129조 제3호)를 보험모집인이 보험계약자나 피보험자에게 특별이익을 제공하는 방식으로 우회하면서 이를 또다시 금지한 것(제98조)과 여신전문금융업법상 대형신용카드가맹점에 대한 수수료율 인하 금지(제18조의3 제1항, 제4항)를 부가통신업자(VAN사)가 대신 대형신용카드가맹점에 리베이트를 지급하는 방식으로 우회하면서 이를 또다시 금지한 것(제19조 제6항)이 있다.

금융AI를 활용하여 금융소비자를 분류하는 모델에 있어 전통적인 계량적 기법에 대비하여 공정성·투명성 관련 증분적 위험(incremental risk)이 무엇인지 (단순한 "블랙박스" 논변을 넘어) 구체적으로 식별할 필요가 있다. 가장 주요한 위험[23]은 일정한 차별금지사유로 차별을 금하는 규제의 준수 및 준수 여부의 검증이 AI의 활용 시 (전통적 계량모델과 비교해서도) 상대적으로 어렵다는 점인데, 그 원인은 다음과 같다.

첫째, 회피가 쉽거나 규제로 인해 도리어 차별이 악화될 수 있기 때문이다. 예컨대 성별을 사유로 한 차별을 금지하는 법이 단지 모델에 투입될 특성값(features)에서 성별변수만 빼는 방식으로 준수된다면(입력중심(input-centric) 접근방식), 성별과 강한 상관관계에 있는 다른 대용변수(proxy)(특정 성별이 많은 대학이나 학부 졸업, 성별에 따라 상이한 구매패턴 등)가 여전히 모델에 투입될 경우 성별변수의 제외가 차별 해소에 도움이 될 수 없다.[24] 많은 경우 고의적인 대용변수 조작(proxy manipulation)이 가능할 뿐 아니라,[25] 의도치 않은 대용변수에 의한 차별금지 우회가 발생할 수 있으며, 특히 핀텍 회사들은 다양한 비금융 데이터를 모아 전례 없이 쉽게 대용변수를 생성할 수도 있다. 더욱이 다른 변수들이 편향되어(biased) 있으면 오히려 성별변수를 빼는 것이 차별을 도리어 악화시킬 수도 있다.[26] 예컨대 청년들이 실제 상환할 의지와 능력이 있음에도 불구하고 단지 기존 신용이력의 부족으로 제2금융권 대출을 많이 이용할 수밖에 없는 상황이라면, 기존 모델에서는 이 상황이 반영되어 청년이 제2금융권을 이용할 시 중장년이 제2금융권을 이용했을 때보다 상환율이 높게 예측될 수 있으나, 연령변수를 뺀 모델에서는 제2금융권 이용 사실만으로 낮은 상환율이 일괄적으로 예측되어 제2금융권 이용률이 높은 청년이 더 불이익해질 수 있다.

둘째, 준수가 어렵기 때문이다. 현재의 차별금지 조항은 차별금지사유("사회적 지위" 등)도, 공정성 판단기준("정당한 사유 없이", "부당하게" 등)도 모호하게 규정한 채 법원의 사후적(*ex post*) 판단에 맡기는 규준(standard)의 형태를 취하는 경우가 많다. 그러나 대량, 다차원의 정형 및 비정형 데이터로 복잡도가 높은 모델을 훈련하고, 경우에 따라서 앙상블(ensemble) 기법에 기해 다수의 약한 학습자들(weak learners)을 결합하여 강한 학습자들(strong learner)을 도출하는 AI모델의 훈련과정에서, 차별금지사유나 공정성 판단기준이 사전적(*ex ante*)으로 엄밀히 정의(formalize)되어 있지 않으면, 공정성 확보를 위해 전처리(preprocessing)를 하기도, 훈련과정에서 제약(constraint) 조건을 두기도, 후처리(postprocessing)를 하기도 어렵다. 즉, 규칙(rule) 수준으로 차별금지사유를 명확히 한정적으로 열거하고 개발자에게 공정성 판단기준을 수리적 지표(metrics)로 제시하지 않으면 알고리듬적 개입(algorithmic intervention)이 사실상 어려워지는 문제가 있다.

셋째, 추적과 평가가 어렵기 때문이다. 기존 선형회귀모델의 경우 모델에 투입되는 각 특성값(설명변수)(X)과 예측($\hat{y}$) 간의 상관관계가 단순한 계수(coefficient)(β)로 표현되는데 반해, 기계학습의 비선형모델(심층학습 모델 포함)의 경우 특성값과 예측의 상관관계가 명확히 표현되지 않을 수 있다. 그리고 앞서 살펴보았듯이 대량, 다차원의 데이터의 투입과 앙상블 기법을 통한 다수 모델의 결합으로 모델 해석가능성(model interpretability)[27]은 더욱 떨어질

23) 이와 더불어 이론적으로 금융소비자에 대한 행태패턴을 포함한 대량의 데이터를 수집하여 이를 AI 모델로 처리하면서 소비자 개개인에 대한 시장세분화(market segmentation) 및 이로 인한 가격차별 가능성이 증대될 수 있으나, 규제로 인해 금융기관들의 가격차별과 관련한 운신의 폭은 좁으므로, 논의를 생략한다.

24) Talia B. Gullis & Jann L. Spiess, *Big Data and Discrimination*, 86 U Chi L Rev 459, 468-471 (2019).

25) Solon Barocas & Andrew D. Selbst, *Big Data's Disparate Impact*, 104 Cal L Rev 671, 712-713 (2016).

26) Gullis & Spiess, *supra* note 24 at 471-473.

27) 모델 해석가능성이란 사용자가 모델의 결과값을 정확하고 효율적으로 예측할 수 있는 정도를 의미한다(Been Kim, Rajiv Khanna & Oluwasanmi O. Koyejo, *Examples are not Enough,*

수 있다. 이로 인해 소송과 규제 집행 과정에서의 추적과 평가가 어려워 차별 입증의 난이도가 높아지고, 이것이 규제 준수의 유인을 더욱 떨어뜨리며 회피의 유인을 높이게 된다. "블랙박스" 논변과 그나마 가장 유관한 측면이다.

결국 AI의 활용 시 그렇지 않은 경우와 동등한 수준의 비차별성을 달성하기 위해서는, 기존 비차별 조항들의 요건 세분화와 명확한 정의가 필요한데, 현 금융소비자 차별금지 조항들은 법관의 사후적 판단과 감독기관의 재량에 맡겨진 모호한 규준에 머물러 있다. 자연법론자들이 환호할 일이나, AI모델의 훈련・검증・시험 단계에서 준수를 가능하게 하려면 이러한 아날로그적 방식으로는 어림도 없다. 비트겐슈타인(Wittgenstein)이 "철학적 탐구"에서 지적한 (일반성에 대한 열망 등으로 인해) 언어가 야기하는 혼란에서 벗어나(§109) 언어를 (현실을 그림처럼 그리는) 명징한 재현(übersichtliche Darstellung)으로 바꾸는 작업(§122)[28] 내지 화이트헤드(Whitehead)와 러셀(Russell)이 "수학원리"에서 제시한 우발적이고 부정확한 현재의 언어를 논리어로 재구성하는 작업[29]을 최소한 흉내라도 내야 한다. 따라서 이하 제2절에서는 규범적 분석을 통해 차별금지 법령 요건의 명징화를 먼저 시도하고, 이에 따라 공정성지표(fairness metrics) 등 공정기계학습(fair ML)의 수리적 방법론이 비차별 조항의 준수를 위해 어떻게 활용될 수 있을지 검토한다. 아울러 제3절에서는 공정성 확보의 수단이라기보다는 완전자동화의 프레임에 함몰된 투명성 내지 설명요구권 규제의 현상에 비판적으로 접근하며 제도 개선을 제안한다.

2. 금융AI에 의한 차별의 금지

(1) 금융소비자 차별금지 법령의 개요

앞서 살펴보았듯이 일정한 차별금지사유를 명시하고 차별을 금지하는 규정들이 있다. 금융소비자보호법 제15조는 "금융상품판매업자등은 금융상품 또는 금융상품자문에 관한 계약을 체결하는 경우 정당한 사유 없이 성별・학력・장애・사회적 신분 등을 이유로 계약조건에 관하여 금융소비자를 부당하게 차별해서는 아니 된다"라 규정하고 있고, 동법 부칙(법률 제17112호, 2020. 3. 24.) 제13조 제5항에 의하여 대부업법 제9조의8이 추가되어 거의 동일한 규정을 하고 있다. 한편, 신용정보법은 개인신용평가회사(CB)가 개인신용평가 시 (1) 성별, 출신지역, 국적 등으로 합리적 이유 없이 차별하거나 (2) 정당한 이유 없이 계열회사와 금융거래 등 상거래 관계를 맺거나 맺으려는 사람의 개인신용평점을 다른 사람의 개인신용평점에 비해 유리하게 산정하는 등 차별적으로 취급하는 행위를 금지한다(제22조의4 제2항 제1호, 제3호, 시행령 제18조의3 제2호).

다음으로 금융투자업법, 보험업법, 여신전문금융업법, 온라인투자연계금융업법은 일정한 차별금지사유의 명시 없이 가격차별(price discrimination)을 제한하는 규정을 두고 있는데,[30] 실무상으로는 (특히

Learn to Criticize! Criticism for interpretability, NIPS 2016, 2288, 2294 (2016)).

28) Ludwig Wittgenstein, Philosophische Untersuchungen §§109, 122 (1953).

29) A.N. Whitehead & Bertrand Russell, Principia Mathematica, Vol I, 1-3 (1963, 2nd Ed).

30) 개별 법령상 차별 금지 조항의 예는 다음과 같다.

- 금융투자업법: 금융투자업자는 투자자 수수료 부과기준 정할 시 정당한 사유 없는 투자자 차별 금지(제58조 제2항). 투자매매업 또는 투자중개업 인가 받아 집합투자증권 판매 영위하는 은행, 보험회사, 종합금융회사는 집합투자증권의 판매업무와 고유 업무 연계한 정당한 사유 없는 고객 차별 금지(제250조 제6항 제2호, 제251조 제2항, 제341조 제1항).
- 보험업법: 보험회사는 보험요율 산출 시 부당한 보험계약자 차별 금지(제129조 제3호). 보험계약 체결・모집 종사자는 체결・모집 시 정당한 이유 없는 장애인 보험가입 거부 금지(제97조 제1항 제10호).
- 여신전문금융업법: 신용카드업자는 신용카드가맹점 수수료율을 공정하고 합리적으로 정하여야 하며 부당한 차별 금지(제18조의3 제1항).
- 온라인투자연계금융업법: 온라인투자연계금융업자(P2P 금융업자)는 이용자 수수료 부과기준 정할 시 정당한 사유 없는 이용

보험요율 산출 시) 상당한 차등은 허용하고 있으나,[31] 기준들이 명확하거나 일관되지는 않다.

상기 법령들에는 상당한 모호함이 있다. 첫째, 금융소비자보호법 제15조가 계약을 체결한 소비자의 공정취급(fair treatment of customers)만을 의미하는지, 아니면 금융상품에 대한 공정접근(fair access to financial services)까지 보장하는 것인지(즉, 차별금지가 포용금융(financial inclusion)까지 포괄하는 개념인지)가 문언상 명확치 않다. 미국의 닫-프랭크법(Dodd-Frank Wall Street Reform and Customer Protection Act of 2010)에 의해 개정된 12 U.S.C. §1이 통화감독청(Office of the Comptroller of the Currency; "OCC")에 감독대상 은행과 연방저축조합의 공정취급 뿐 아니라 공정접근을 감독할 권한을 명시적으로 부여하는 것[32]과 다르다. 금융소비자보호법상 "금융소비자"가 계약상대방 뿐 아니라 권유 대상자나 청약자도 포함하는 것(제2조 제8호)을 고려하여 "계약조건"을 계약 체결 또는 거절 여부를 포함하는 것으로 확대해석하면 공정접근의 문제를 포괄할 여지가 없진 않다. 그러나 "계약을 체결하는 경우" "계약조건에 관하여" 차별을 금지한 것은 대출이나 보험가입 거절 등으로 계약이 아예 "체결"되지 않아 "계약조건(terms and conditions; Vertragsbedingung)"이 존재하지 않는 경우(공정접근의 문제)는 포괄하지 않는 것으로 해석하는 것이 문언의 통상적 의미에 충실히 해석하는 일반적 법해석 원칙에는 더 부합한다.[33] 다만 정책적으로는 법이 공정접근은 보장하지 않고 공정취급만 강제할 경우, 공정취급 의무로 인해 리스크 프리미엄(risk premium)을 실제 관측된 신용리스크만큼 부과하지 못하는 그룹(protected group)에 대해서는 금융기관이 접근을 더 차단하는 선택(대출 또는 보험가입 거절 등)을 하게 되어 공정접근은 오히려 악화될 수밖에 없다. 중장기적인 관점에서 이하에서는 공정취급 뿐 아니라 공정접근의 문제도 함께 검토한다.

둘째, 차별금지사유(prohibited bases/factors 또는 protected/sensitive characteristics)의 명징한 정의가 필요하다. 차별금지사유는 헌법상 "성별·종교 또는 사회적 신분"(제11조 제1항), 금융소비자보

자 차별 금지(제11조 제4항). 차입자 정보 제공, 투자자 모집, 원리금 상환 등 업무수행 시 특정 이용자 부당 우대 또는 차별 금지(제12조 제8항).

31) 보험업의 경우 생명보험, 손해보험 보험사는 각 위험요소가 위험보장에 영향을 미치는 정도에 따라 보험가입을 거절하거나 보험가입금액 한도 제한, 일부 보장 제외, 보험금 삭감, 보험료 할인·할증 등 조건부 인수를 할 수 있으며 보험종목별로 그 기준이 되는 계약인수지침을 마련하여 사용한다(보험업감독업무시행세칙 제5-13조, [별표 14] 표준사업방법서 생명보험 제9조 제1항, 손해보험 제10조 제1항). 금융투자업의 경우 투자중개업자의 협의 위탁매매수수료는 영업기여도(자산(수익)규모), 고객 등급, 고객 중요도(지점장평가)에 따라 차등할 수 있다(금융투자협회 「금융투자회사의 영업 및 업무에 관한 규정」제2-62조, 동 시행세칙 제6조, 별지 제3호 "수수료 부과 기준"). 여신전문금융업의 경우 영세·중소·특수·대형 가맹점별로 신용카드 평균결제금액별 차등을 할 수 있다(여신전문금융업법 시행규칙 제3조 제1호 다목, 여신금융협회 신용카드 가맹점 표준약관 제9조 제11호, 가맹점 매출거래정보 통합조회서비스(https://www.cardsales.or.kr/)). 참고로, 은행업의 경우 1991~1997년 4단계를 거쳐 금리자유화가 이루어졌다.

32) "There is established in the Department of the Treasury a bureau to be known as the 'Office of the Comptroller of the Currency' which is charged with assuring the safety and soundness of, and compliance with laws and regulations, fair access to financial services, and fair treatment of customers by, the institutions and other persons subject to its jurisdiction" (12 U.S.C. §1(a)).

33) "법해석의 목표는 어디까지나 법적 안정성을 저해하지 않는 범위 내에서 구체적 타당성을 찾는 데 두어야 한다. 그리고 그 과정에서 가능한 한 법률에 사용된 문언의 통상적인 의미에 충실하게 해석하는 것을 원칙으로 하고, 나아가 법률의 입법 취지와 목적, 그 제·개정 연혁, 법질서 전체와의 조화, 다른 법령과의 관계 등을 고려하는 체계적·논리적 해석방법을 추가적으로 동원함으로써, 위와 같은 법해석의 요청에 부응하는 타당한 해석이 되도록 하여야 한다 … '임대차 조건'이라는 문언은 통상적으로 임대차 계약관계에서 계약 내용을 이루는 사항을 의미하는데, 이러한 계약 내용으로는 일반적으로 임차인이 사용대가로 지급하는 차임과 임대차의 기간이 핵심적 사항이 되는 점 …" (대법원 2010. 12. 23. 선고 2010다81254 판결). 공정거래법 제23조 제1항 제1호가 "부당하게 거래를 거절하거나 거래의 상대방을 차별하여 취급하는 행위"를 불공정거래행위로 규정하는 것도 문언해석상 거래 상대방 차별에는 거래 거절이 포괄되기 어렵기 때문이다.

호법상 "성별·학력·장애·사회적 신분 등"(제15조), 신용정보법상 "성별, 출신지역, 국적 등"(제22조의4 제2항 제1호)으로 규정되어 있다.[34] 성별·종교는 헌법이 개별적으로 특정한 차별금지사유라는 점에서 이들을 각 법률에서 차별금지사유로 명시하는 것은 문제가 없다(특히 성별 차별 금지는 헌법 제34조 제3항이 추가적인 근거가 된다). 장애의 경우 헌법이 장애인에 대한 법률이 정하는 바에 의한 국가의 보호를 규정하고(제34조 제5항) 이에 따라 장애인차별금지법이 대사인효가 있는 차별금지법으로 제정되어 있으며 동법이 금융에 대한 특칙까지 두고 있으니(제17조) 차별금지사유로 규정해야 전체 법체계에 부합한다. 문제는 학력·출신지역·국적 등 기타 사유와 "사회적 신분"이라는 일반 사유이다. 특히 (1) 미국의 공정신용기회법(Equal Credit Opportunity Act, 15 U.S.C. §1691 et seq.; "ECOA")[35]이 신용제공자(creditor)의 신용거래 청약자에 대한 차별금지사유로 인종·피부색, 종교, 출신국가, 성별, 혼인상태, 연령(신청인이 행위능력이 있다는 전제), 공공부조 수령 사실, 소비자신용보호법(CCPA)상 권리행사 사실을 열거하고 있는 점(§1691(a)), (2) 미국의 민권법(Civil Rights Act of 1968)의 제8-9장인 공정주거법(Fair Housing Act, 42 U.S.C. §3601 et seq.; "FHAct")[36]이 담보대출자(mortgage lender)에도 적용되면서 차별금지사유로 인종·피부색, 종교, 출신국가, 성별, 가족관계, 장애를 열거하고 있는 점(§3604), (3) 흔히 오바마케어법이라 불리는 환자보호·의료비 적정화법(Patient Protection and Affordable Care Act, 42 U.S.C. §18001 et seq.; "ACA")이 연방재정지원을 받는 건강보험 프로그램 및 활동에 적용되면서 차별금지사유로 인종, 피부색, 출신국가, 성별, 연령, 장애[37]를 열거하고 있는 점(§18116)과 비교하여 살펴본다.

- 학력: 학력이란 용어를 쓰니 학벌 같은 부정적 연상만 떠오를 수 있으나, 법에 차별금지사유로 명기하는 이상 교육수준의 고려를 금하는 것이 된다. 미국에서의 연구이나 교육수준이 높아지면 (투자소득 및 자산소유도로 측정하는) 금융시장 참여도가 높아지고, 파산·압류·채무불이행 비율이 극적으로 줄어든다는 사실이 실증된 바 있다.[38] 외국에서 교육수준이 금융시장의 차별금지사유로 규정된 사례는 없다고 추정된다. 우리만 유별나진 계기 중 하나가 감사원의 2012. 7. 23.자 금융감독당국 감사 결과이다.[39] 감사원은 금융감독원이 2008. 4. 28. 신한은행의 Basel II에 따른 신용리스크 내부등급법 사용을 승인하는 과정에서 학력을 직업·급여 외 별도 항목으로 평가하여 신용평점에 차등(고졸 이하 13점~석·박사 54점)을 둔 신용평가시스템(CSS)을 검토·승인한 것이 동 은행의 학력 차별에

34) 국가인권위원회법은 성별, 종교, 장애, 나이, 사회적 신분, 출신지역, 출신 국가, 출신 민족, 용모 등 신체 조건, 혼인 여부, 임신 또는 출산, 가족 형태 또는 가족 상황, 인종, 피부색, 사상 또는 정치적 의견, 형의 효력이 실효된 전과, 성적 지향, 학력, 병력 등 19개 차별금지사유를 열거(제2조 제3호)하고 있으나, 동법상 평등권 침해의 차별행위 발생 시 원칙적으로 국가인권위원회의 구제조치 등의 권고 사유(제44조)일 뿐 곧바로 민사책임 또는 행정제재로 이어지지 않음에 유의할 필요가 있다.

35) ECOA는 연준위(FRB)의 Regulation B (12 CFR §202), 소비자금융보호국(CFPB)의 Regulation B (12 CFR §1002)에도 그대로 반영되어 있다.

36) FHAct는 주택도시개발부(HUD)의 행정규칙인 24 CFR §100에 근거하여 집행된다.

37) 보건복지부(HHS) 행정규칙인 45 CFR §92가 차별금지사유를 구체화하나, 세부 범위에는 정치적 논쟁이 있었다. 오바마 정부가 2016. 5 18. 제정 시 "임신, 거짓 임신, 임신 중단, 이로부터의 회복, 출산 또는 관련된 의학적 상태, 성역할 고정관념, 성적 취향(pregnancy, false pregnancy, termination of pregnancy, or recovery therefrom, childbirth or related medical conditions, sex stereotyping, and gender identity)"에 기한 차별도 성별에 근거한 차별의 하나로 규정했으나(81 FR 31375-473), 트럼프 정부가 2020. 6. 19.자 개정으로 성별을 남녀 구분으로 제한했다(85 FR 37160-248).

38) Shawn Cole, Anna Paulson & Gauri K. Shastry, *Smart Money? The Effect of Education on Financial Outcomes*, 27(7) Rev Financ Stud 2022 (2014).

39) 고학수/정해빈/박도현, "인공지능과 차별", 저스티스 제171호(2019. 4.), 한국법학원, 199-277면.

대한 지도·감독 부적정이라 지적하였다.[40] 이에 대응한 감독당국의 행정지도에 따라 전국은행연합회는 2012. 1. 30.「불합리한 차별행위 방지 모범규준」을 마련하여 "은행이용자의 성별, 종교, 장애, 나이, 출신지역, 출신국가, 용모 등 신체조건, 혼인여부, 사상, 성적 지향, 학력 등을 이유로 특정 은행이용자를 우대·배제·구별하거나 불리하게 대우하는 행위"를 "차별행위"로 규정(제2조 제4호)하면서 은행의 신용평가, 금융상품 권유, 계약체결 및 유지, 거래조건에서의 차별행위를 금지하였다(제3조).[41] 전국은행연합회는 특히 "학력은 통계적으로 의미 있는 신용평가지표로서 외국에서도 차별금지사유로 규제하고 있지 않으나, 금번에 은행권은 학력차별근절에 대한 사회적 요구 등을 고려하여 차별사유에 포함"이라 밝혔다.[42] 동 모범규준은 2014. 11. 12. 감독당국의 행정지도 정비 과정에서 폐지된 것으로 보인다.[43] 그러나 학력은 금융소비자보호법 제15조에 은행 뿐 아니라 금융 전체를 아우르는 차별금지사유로 명기되어 부활하였다.

- 국적: 개인신용평가에 있어 국적의 고려를 금하는 것도 매우 이례적인 것으로 보인다. 미국의 ECOA와 FHAct 상의 차별금지사유인 "출신국가(national origin)"란 미국 시민권을 전제로 본인 또는 선조가 태어난 국가를 의미하는 혈통 개념이지, 국적을 의미하는 것이 아니다. 실증적 연구는 필요하나 외국에 생활 거점이 있을 가능성이 높고 내국인과 다른 경제적, 문화적 환경에 놓여 있는 외국인에게 내국인과 동등한 신용평가를 하도록 강제하는 것은 지나친 이상론으로 보인다.
- 출신지역: 국적과 달리 생래적이고 위험도와 관련도 낮을 것으로 추정되는 출신지역을 사유로 한 차별을 금하는 것은 합리적이라고 생각된다.
- 사회적 신분: 헌법재판소는 헌법 제11조 제1항의 차별금지사유인 "사회적 신분"이란 "사회적 신분이란 사회에서 장기간 점하는 지위로서 일정한 사회적 평가를 수반하는 것"을 의미한다며 전과도 포함한다고 보았다.[44] 하급심 법원들은 근로기준법 제6조의 차별금지사유인 "사회적 신분"을 "선천적 신분뿐만 아니라 자기 의사에 의해서도 피할 수 없는 후천적 신분도 포함"[45]된다거나 "사업장 내에서 근로자 자신의 의사나 능력발휘에 의해서 회피할 수 없는 사회적 분류"라 해석하나,[46] 구체적인 사안에서 일관된 기준을 제시하지는 못하고 있다.[47] 현 금융소비자보호법 제15조와 관련해서는 특히 본인의 의사와 능력발휘에 의하여 취득한 지위로서 위험도와 유의미한 상관관계가 있는 후천적 신분은 사회적 신분으로 간주하는 것에 신중할 필요가 있다. 지금까지 특히 보험감독 실무상 이하 사유가 차별금지사유로 취급된 바 있으나, "사회적 신분" 해당 여부는 특히 다른 금융업 영역과의 관계에서 신중히 검토되어야 한다.
  - 직업·직종: 보험업감독업무시행세칙 첨부 표준사업방법서에 2020. 7. 3. "역선택 방지 등 합리

40) 감사원, 2012. 7., "감사결과 보고서 - 금융권역별 감독실태 II (은행, 신용카드, 보험 권역 등을 중심으로)," 31-32면, https://www.bai.go.kr/bai/cop/bbs/detailBoardArticle.do?bbsId=BBSMSTR_100000000009&nttId=1387.

41) 전국은행연합회, 2012. 10. 31., "은행권,「불합리한 차별행위 방지 모범규준」 제정·시행," https://www.kfb.or.kr/news/info_news_view.php?idx=461.

42) 앞의 보도자료.

43) 아시아투데이, 2014. 11. 16., "차별금지가 불합리한 규제? … 금융당국, 고객차별금지 모범규준 폐지."

44) 헌법재판소 1995. 2. 23. 93헌바43 결정.

45) 서울고등법원 2017. 11. 24. 선고 2016다2070186 판결 (대법원 2018. 3. 29. 선고 2017다293131 판결로 심리불속행 확정).

46) 대전고등법원 2015. 11. 26. 선고 2014나11589 판결 (대법원 2019. 12. 24. 선고 2015다254873 판결로 상고기각 확정).

47) 사회적 신분에 무기계약직은 포함되지만(위 대전고등법원 2014나11589 판결 및 서울중앙지방법원 2018. 6. 14. 선고 2017가합507736 판결 (확정)), 비정규직(서울고등법원 2012. 12. 7. 선고 2012나39631 판결 (대법원 2015. 10. 29. 선고 2013다1061 판결로 상고기각 확정)), 동일 고용형태(무기계약직) 내 직종 중 하나(위 서울고등법원 2016나2070186 판결), 회사를 상대로 소송 등을 제기한 사실(대구지방법원 2019. 7. 10. 선고 2018나319922 판결 (확정))은 포함되지 않는다고 본다.

적인 사유 없이 특정 직업 또는 직종에 종사한다는 사실만으로 보험가입을 거절하지 않습니다"라는 규정이 추가되었다(세칙 제5-13조, [별표 14] 생명보험 제9조 제3항, 손해보험 제10조 제2항). 개정 당시 유독 소방관·군인이 강조되었는데, 사회적으로 존중받는 소수 특이례가 모든 직업·직종의 일률적 차등 금지에 대한 근거가 될 수는 없다. 국가가 소방관·군인을 위해 마땅히 할 일을 보험계약자들로부터 세금을 걷듯이 떠넘기는 셈인데, 예산을 배정하여 위험수당을 인상하고 소방·군인공제제도를 확충하면 다른 고위험 직업에 대한 누수 없이 더 제대로 지원할 수 있다.

- 거주지역: 출신지역과 달리 비생래적이고 위험도와 관련성도 큰 거주지역을 "사회적 신분"이라 단정하기 어렵다. 미국에서는 거주지역별로 여신과 보험에 차등을 두는 것을 흔히 각각 레드라이닝(redlining)과 우편번호 요율산정(ZIP code rating)이라 칭하는데, 미국은 거주지역이 인종별로 나눠져(racially segregated) 있다 보니 특히 레드라이닝의 경우 FHAct 해석상 인종차별로 간주하는 경우가 있고, 공동체재투자법(Community Reinvestment Act, 12 U.S.C. §2901; "CRA")에 따라 예금보험제도 적용을 받는 은행들이 모든 지역에서 서비스를 제공하도록 독려하나, 우리의 상황이 같다고 하기 어렵다. 금융소비자보호법 제정 과정에서 한국씨티은행의 지점 폐쇄를 제15조의 차별금지사유에 "지역"을 추가하는 방식으로 규제하자는 의견도 제시되었으나[48] 영업망의 확충 문제는 차별금지보다는 인가조건 준수 여부 등의 문제로 접근해야 한다. 보험의 경우 교통사고 발생률의 지역별 격차가 큰데도 자동차보험의 지역별 요율 차등화가 이루어지지 못하여 논란이 지속되고 있다.[49]
- 장기기증: 장기이식법이 "누구든지 장기등 기증을 이유로 장기등기증자를 차별대우하여서는 아니 된다"(제3조 제2항)라 규정하고 있어 보험감독 실무상 차별금지사유로 취급해왔는데,[50] 위 법령과의 조화로운 해석상 "사회적 신분"으로도 인정할 수 있을 것이다.

셋째, 위 논의들을 종합하여 합리적 차별의 판단기준을 구체화해야 한다. 대법원이 공정거래법상 일반적인 차별적 취급행위의 요건인 "부당하게"는 공정거래위원회에 입증책임이 있으나 계열회사를 위한 차별적 취급행위의 요건인 "정당한 이유 없이"는 행위자가 입증책임이 있다고 구분하는 상황에서,[51] 금융소비자보호법 제15조가 "정당한 사유"와 "부당하게" 양자를 중복적으로 요건에 넣은 것은 타당해 보이지는 않는다. 다만, 현 문구상으로는 대체로 원고나 감독당국이 일응의 부당성을 입증하면 금융상품판매업자등이 합리적 근거에 의한 차별임을 항변하는 구조로 해석할 수밖에 없다.

헌법상 평등원칙과 관련하여 헌법재판소는 합리적 근거에 의한 차별을 허용하는 상대적 평등으로 해석하면서, 합리적 근거 여부는 인간의 존엄성 존중과 정당한 입법목적 달성을 위해 필요, 적정한지 여부를 기준으로 판단하여야 한다고 보고 있다.[52] 헌법재판소는 1999년의 제대군인 가산점 위헌결정 이후로는 (1) 성별 등 엄격심사를 적용해야 할 사안에서는 입법형성권이 축소되어 비례성 원칙이 적용되나 (2) 기타 사안에서는 입법형성권을 존중하여 자의금지 원칙에 따르도록 하고 있다.[53] 이는 차별금지사유별로 합리적

48) 전성인, "금융소비자보호법 제정과 관련한 진술," 금융소비자 보호에 관한 공청회 (2017. 9.), 국회 정무위원회, 88-9면.

49) 정중영/강준규, "자동차보험 지역별 요율 차등화에 관한 연구," Journal of the Korean Data Analysis Society 제8권 제1호 (2006. 8.), 한국자료분석학회, 363-373면.

50) 금융감독원, "장애인 및 장기기증자에 대한 보험가입 차별금지 독려" (2009. 5. 20.), https://www.fss.or.kr/fss/kr/promo/bodobbs_view.jsp?seqno=13584 및 금융감독원, "장애인 및 장기기증자에 대한 보험가입 차별금지 홍보 강화" (2010. 5. 19.), https://www.fss.or.kr/fss/kr/promo/bodobbs_view.jsp?seqno=14377 등.

51) 대법원 2001. 12. 11. 선고 2000두833 판결.

52) 헌법재판소 1998. 9. 30. 98헌가7 결정.

차별 여부를 달리 판단하는 미국 연방대법원의 단계별 기준(sliding-tier standard)[54]을 일부 받아들여 판단기준의 명확성을 제고한 것으로 볼 수 있다. 그러나 제대군인의 공무원 채용 시 5% 또는 3% 가산은 위헌으로,[55] 국가유공자 또는 유족의 국가기관 채용 시 10% 가산은 합헌으로,[56] 국가유공자 가족의 국공립학교 채용 시 10% 가산은 헌법불합치로[57] 판시하는 등 사전적(ex ante) 지침으로 삼을만한 일관된 기준을 제시하지는 못하고 그때그때의(ad hoc) 사후평가에 그치고 있는 실정이다.

미국의 ECOA나 FHAct 상의 공정대출(fair lending)의 판단기준을 참고하여 구체화의 실마리를 얻을 수 있다. 미국 법원들은 차별적 대우(disparate treatment) 또는 차별적 효과(disparate impact)가 입증되면 차별을 인정한다. 차별적 대우는 차별적 의도나 편견을 가지고 공개적으로 차별을 가하는 공공연한 증거(overt evidence) 뿐 아니라 그러한 의도나 편견 없이도 차별금지사유에 기초하여 다르게 대우한다는 비교적 증거(comparative evidence)(정황증거(circumstantial evidence))에 의해 입증될 수 있다.[58] 차별적 효과란 형식적으로는 중립적으로 대우하더라도 결과적으로 보호대상 집단을 불균형하게 배제하거나 불이익을 주는 경우를 의미한다. 기존에 노동관계(민권법 제7장)에 적용되었던 차별적 영향이 공정대출의 판단에 있어서도 적용되는지 여부에 대해서 여러 하급심 판결들과 논란이 있었는데, 미국연방대법원이 2015년 차별적 영향이 FHAct 상 공정대출의 판단기준이 될 수 있다고 5:4로 결론 내렸다.[59] 금융AI의 맥락에서 공공연한 증거에 의한 차별적 대우는 매우 드물 것이다. 정황증거에 의한 차별적 대우는 차별금지사유 변수를 모델 훈련에 투입(input)하는지 여부에 따라 판단될 것이므로 이를 입력중심(input-centric 또는 input-focused) 접근으로 볼 수 있다.[60] 반면 차별적 효과는 훈련된 모델의 출력(output)에 있어 결과적으로 차등이 이루어졌는지를 보는 것이므로, 출력중심(output-centric 또는 output-focused) 접근에 해당한다.[61] 다른 한편, 앞서 살펴본 바와 같이 닫-프랭크법은 통화감독청(OCC)에 감독 대상 은행과 연방저축조합의 공정취급과 공정접근을 감독할 권한을 부여하는데, 이 수권 하에 OCC는 2020. 11. 25. 연방관보(Federal Register)에 "금융 서비스에 대한 동등접근(Fair Access to Financial Services)"이라는 행정규칙안을 입법예고하였다(12 CFR §55).[62] 그 핵심은 합리적인 개별적 신용평가를 한 경우에는 공정접근 의무가 없고 그 이외의 경우 공정접근을 보장해야 한다는 것이나,[63] 신중한 검토의 필요성이 제기되면서 OCC

53) 헌법재판소 1999. 12. 23. 98헌마363 결정.

54) 미 연방대법원 판례상 분류(classification)가 수정 제14조의 평등조항(equal protection clause)에 합치되는지 여부는 (1) 인종 등 "엄격심사(strict scrutiny)"가 필요한 "의심(suspect)" 분류(Korematsu v. United States, 323 U.S. 214, 216 (1944) 등), (2) 성별 등 "중간 단계 심사(intermediate scrutiny)"가 필요한 "중간(intermediate)" 분류(Craig v. Boren, 429 U.S. 190, 197 (1976) 등), (3) 연령, 부 등 "최소합리성 기준(minimum rationality standards)"이 적용되는 "기타" 분류(Massachusetts Bd. of Retirement v. Murgia, 427 U.S. 307 (1976) 등)로 구분하여 심사한다.

55) 위 헌법재판소 98헌마363 결정.

56) 헌법재판소 2001. 2. 22. 2000헌마25 결정.

57) 헌법재판소 2006. 2. 23. 2004헌마675, 981, 1022 결정.

58) McDonnell Douglas v. Green, 411 U.S. 792 (1973).

59) Texas Dept. of Housing and Community Affairs v. Inclusive Communities Project, Inc., 576 U.S. 519 (2015)

60) Gullis & Spiess, *supra* note 24 at 472.

61) Id at 472.

62) OCC, Notice of Proposed Rulemaking: Fair Access to Financial Services, 12 CFR §55, https://www.occ.gov/news-issuances/federal-register/2020/85fr75261.pdf.

63) 구체적으로, 대형 은행 등(총자산 1천억 달러 이상일 경우 추정)(covered bank)은 (1) 자신이 제공하는 각각의 금융상품을 지리적 시장 내의 모든 사람들에게 비례적으로 동일한 조건으로 제공해야 하고, (2) 자신이 제공하는 금융상품을 누구에게도 거절해서는 아니 되나, 그 사람이 미리 설정한 양적이고 불편부당한 리스크 기반 기준을 맞추는 데 실패했다는 계량화되고 문서화된 근거가 있으면 거절 가능하며, (3) 특수관계자를 위해 진입 또는 경쟁을 제한하는 방식으로 거절하는 것은 금지되고, (4) 타인과 공동(coordination)으로 금융상품을 거절해서는 아니 된다는 내용이다(OCC, 위 행정규칙).

는 2021. 1. 28. 이 규정의 공포를 보류하고 차기 바이든 정부에 넘겼다.

특히 보험의 경우 성별에 따른 보험료율 차등이 가능할지가 쟁점이다. 생명보험 보험료 산출과 관련하여 보험업감독업무시행세칙 첨부 표준사업방법서가 "보험료율은 피보험자(보험대상자)를 기준으로 남자 또는 여자의 성별 구분에 따라 각각 별도로 산출하여 적용한다"라 규정하여(세칙 제5-13조, [별표 14] 생명보험 제17조) 성별에 따른 차등을 허용해 왔다. 40대 피보험자가 2018. 1. 현재 가입금액 1억원 상품에 가입하는 것을 기준으로 생명보험 보험료는 평균수명이 짧고 위험노출이 큰 남성이 여성보다 18.7% 비쌌다고 한다.[64] 유럽의 경우 재화 및 용역 동등취급 지침(Equal Treatment in Goods and Servives Directive 2004; Directive 2004/113/EC)이 성별별 보험료와 보험급부의 비례적 차등을 명시적으로 허용하였으나(§5(2)) 유럽연합법원(ECJ)은 2011년 이 조항이 유럽연합 기본권 헌장(Charter of Fundamental Rights of the European Union; "CFR") 제21, 23조에 위배된다는 이유로 무효화했다.[65] 전반적으로, 성별을 금융소비자보호법상 차별금지사유로 명기한 이상 유독 보험만 예외로 두는 것은 난점이 있다.

(2) 금융 AI의 활용 관련 차별 금지 조항의 적용

1) 차별금지사유별 부당성과 정당한 사유의 판단기준의 명징화

법은 차별금지사유별로 부당성이나 정당한 사유를 달리 판단할 수 있는지 명시하지 않으나, 앞서 살펴본 헌법재판소의 1999년 제대군인 가산점 위헌결정은 단계별 기준(sliding-tier standard)이 가능함을 보여주고 있으므로, 사유별로 일응의 "부당한" 차별(prima facie case of discrimination)에 해당하는지 여부의 기준(단, 금융상품판매업자 등의 "정당한 사유" 항변 가능)을 다음과 같이 나누어 살펴볼 수 있다.

- 성별: 헌법이 특정하여 보호하고 이로 인해 헌법재판소도 엄격심사(strict scrutiny)를 적용하고 있으므로, 부당성 판단과 관련해서는 차별적 대우 기준이나 입력중심 접근에 머무를 수는 없고, 차별적 영향(disparate impact) 기준 내지 출력중심(output-centric) 접근을 병행해야 한다.
- 장애: 전체 법체계와의 조화로운 해석상 차별적 영향을 기준으로 한 출력중심 접근도 필요하다. 장애인차별금지법은 제4조 제1항이 "장애인을 장애를 사유로 정당한 사유 없이 제한・배제・분리・거부 등에 의하여 불리하게 대하는 경우"(1호)(즉, 차별적 대우) 뿐 아니라 "장애인에 대하여 형식상으로는 제한・배제・분리・거부 등에 의하여 불리하게 대하지 아니하지만 정당한 사유 없이 장애를 고려하지 아니하는 기준을 적용함으로써 장애인에게 불리한 결과를 초래하는 경우"(2호)(즉, 차별적 영향)를 명시적으로 차별행위로 간주하기 때문이다.[66]
- 학력, 국적: 앞서 살펴본 여러 난점을 고려하면, 차별적 대우(disparate treatment)에 해당하는 경우, 즉 해당 변수를 직접 AI모델에 투입한 경우에만 일응의 부당한 차별에 해당한다고 보아야 한다. 특히 학력과 관련하여, 2012년 신한은행 사건에서의 감사원도 "통상 개인별 학력 차이는 직업이나 급여 등에 영향을 미치고 있는데도 학력을 직업이나 급여 이외에 별도 항목으로 평가하여 학력별로 신용평점에 차등"[67]한 것만을 문제 삼았다. 다시 말해,

64) 헤럴드경제, "남성 보험료, 여성보다 16.3% 비싸다" (2018. 1. 10.).

65) Association belge des Consommateurs Test-Achats ASBL v Conseil des Ministres (2011) C-236/09.

66) 다만, 제17조가 금융상품 및 서비스의 제공자는 정당한 사유 없이 장애인을 "제한・배제・분리・거부"하여서는 아니 된다라고만 규정하고 있어 차별적 대우만을 규정한 것이 아닌지 의문이 있을 수 있으나, 장애인차별금지법은 우리 법에서는 (노동차별금지 법제를 제외하면) 상당히 예외적으로 민간영역에 곧바로 적용되는 구조이므로 일반적으로 차별적 영향이 금지되어 있는 이상 제17조의 해석에 있어서도 차별적 영향도 고려해야 한다고 볼 수밖에 없다.

67) 감사원, 위 감사결과 보고서 (주 40).

학력을 변수에서 빼기만 하면 학력과 상관관계가 큰 직업・급여가 대용변수로 결과를 좌우하는 것은 상관없다는 것이다. 금융거래이력・소득・자산・직장 등 신용도에 대한 설명력이 큰 변수들이 대부분 교육수준의 대용변수나 다름없어 이러한 입력중심(input-centric) 접근은 불가피하다.

- 기타 "사회적 지위": 전형적으로는 인종, 종교, 장기기증 등이 여기에 해당할 수 있을 터인데, 차별적 대우 뿐 아니라 차별적 영향 기준도 적용할 수 있을지는 실증에 기한 사유별 개별 검토가 필요하다.

2) 공정성 지표(fairness metrics)에 의한 AI 모델의 차별적 영향 측정[68]

요컨대, 성별과 장애 등 변수를 AI모델에 투입되는 특성값 벡터(X)로부터 제외함으로써 차별적 대우로부터 제외되는 것을 전제로, 차별적 영향 여부에 따라 출력중심(output-centric)으로 금융소비자보호법 제15조 등 위반 여부를 판단해야 한다. 그 구체적인 기준과 관련해서는 최근 공정기계학습(fair ML) 논의 과정에서 제안되고 있는 공정성 지표를 참조할 수 있다.

공정성과 관련된 대표적 오해 중 하나는 공정성이 늘 정확성과 상충(tradeoff) 관계에 있다는 것이다. 그 오해는 공정성 여부를 엄밀한 분석 없이 어림짐작하여 판단하는 악습, 그렇지 않더라도 (아래서 살펴보는 "독립성" 지표와 같이) 할당과 안배의 틀에 갇힘에 기인한다. 자질(merit)이 적은 사람이 척도의 빈틈을 잘 타 좋은 대우를 누린다면 횡재(windfall)일 뿐 공정이 아니다. 결국 공정성을 논하기에 앞서 정확성 지표(accuracy metrics)를 이해해야 한다. 기계학습 개발자들이 통상 쓰는 정확성 지표란 데이터셋을 훈련셋(train set)과 테스트셋(test set)으로 나눈(split) 후 훈련셋으로 먼저 모델을 훈련하고 이렇게 훈련된 모델을 테스트셋으로 테스트한 결과 전체 예측에서 정확히 예측된 샘플이 얼마나 되는지 비율을 따지는 것이다. 예컨대 테스트셋으로 분리된 100명의 과거 신용평가 대상자 중 신용양호자로 예측되는 70명 중 50명이 실제 갚았고, 신용불량자로 예측되는 30명 중 20명이 실제 갚지 않았다면, 100명 중 70명이 정확히 분류되었으니, 예측정확도(predictive accuracy)는 70%이다.

이러한 계산법은 위음성(False Negative; "FN" = 제1종 오류(Type 1 error))(위 예에서 신용양호가 양성이라면, 실제 갚은 60명 중 신용불량자로 잘못 분류한 10명)과 위양성(False Positive; "FP" = 제2종 오류(Type 2 error))(위 예에서 실제 안 갚은 40명 중 신용양호자로 잘못 분류한 20명)에 1:1의 같은 비중을 두는 것을 전제한 것이다. 그러나 사람을 분류하는 모델은 통상 위음성과 위양성의 비중이 다르다. 예컨대 의학진단에서는 통상 FN의 최소화가 중요하나(병에 걸렸는데 건강하다고 진단하면 치료시기를 놓치거나 타인에게 감염 우려), 양성 진단 시 정신적 충격이 수반되는 질병(불치병, 성전파성 질환)은 FP의 최소화도 중요하다. 채용심사나 대출심사 등 혜택을 받을 대상자를 분류할 경우 FN(자질 있는데 떨어지는 사람)의 최소화가 중요하고, 형사사법이나 부정거래 탐지 등 제재를 받을 대상자를 분류할 경우 FP(무고한 사람)의 최소화가 중요하다. 이를 판단하기 위해 위음성(FN), 위양성(FP), 진양성(True Positive; "TP"), 진음성(True Negative; "TN")을 다음과 같이 표로 정리한 것이 혼동행렬(confusion matrix)이다.

위양성 또는 위음성 중 하나에만 초점을 두어 정확도를 측정하는 방법에는 크게 두 가지가 있다. 첫 번째는 실제(ground-truth)을 기준으로, 실제 양성인 집단과 실제 음성인 집단 중 각각 제대로 분류된 비율이 어느 정도인지를 측정하는 것인데, 전자를 민감도

68) 본 항목에서 설명하는 공정성 지표에 대한 논의를 위해 다음 자료들을 전체적으로 참조하였다.

- Solon Barocas, Moritz Hardt & Arvind Narayanan, *Fairness and Machine Learning: Classification* (2019), https://fairmlbook.org/pdf/classification.pdf.
- Sahil Verma & Julia Rubin, *Fairness Definitions Explained* (2019), https://fairware.cs.umass.edu/papers/Verma.pdf.

		예측 (predicted; $\hat{y}$)		
		양성 (positive)	음성 (negative)	
실제 (actual; y)	양성 (positive)	진양성 (TP)	위음성 (FN)	sensitivity (recall) = TP/(TP+FN)
	음성 (negative)	위양성 (FP)	진음성 (TN)	specificity = TN/(TN+FP)
		precision = TP/(TP+FP)	NPV = TN/(TN+FN)	

(sensitivity)(또는 재현율(recall)), 후자를 특이도(specificity)라고 하며 위 표에서와 같이 계산한다. 두 번째는 예측(prediction)(또는 분류(classification))을 기준으로, 양성으로 예측된 집단과 음성으로 예측된 집단 중 각각 제대로 분류된 비율이 어느 정도인지를 측정하는 것인데, 전자를 정밀도(precision)(또는 양성예측치(positive predictive value; "PPV")), 후자를 음성예측치(negative predictive value; "NPV")라고 한다.

이제 공정성 지표로 넘어가면, 크게 독립성(independence), 분리성(separation), 충분성(sufficiency)의 세 기준이 제시된다.[69] X를 Y(={+, -})로 분류함에 있어, A(={a, b})가 차별금지사유로 규정된 민감속성(남・여, 장애의 유・무)이고, R(={+, -})이 분류모델의 예측(대출 가능, 보험가입 가능 여부)이라 하자. 그러면 각 기준은 다음과 같이 정의된다.

- 독립성($R \perp A$): 정확성을 따지지 않고 결과적 평등과 안배를 요하는 기준으로서 다음과 같이 표현된다.[70]

$$\boldsymbol{P(R = +|A = a) = P(R = +|A = b), \forall a, b \in A}$$

대출심사의 예에서는, 남성의 대출 확률 = 여성의 대출 확률(따라서 남성의 대출거절 확률 = 여성의 대출 거절 확률)을 의미한다.

- 분리성($R \perp A | Y$): 실재(ground-truth)를 기준으로 한 정확도(sensitivity와 specificity)가 남녀간, 장애인・비장애인간 동일하도록 하는 것이다.[71]

$$\boldsymbol{P(R = +|Y = y, A = a) = P(R = +|Y = y, A = b),}$$

$$\boldsymbol{\forall a, b \in A, y \in \{+, -\}}$$

대출심사의 예에서는, 결국 나중에 돈을 갚을 만한 남성과 나중에 돈을 갚을 만한 여성은 대출이 승인될 확률(sensitivity)이 같아야 하며, 나중에 돈을 안 갚을 만한 남성과 나중에 돈을 안 갚을 만한 여성은 대출이 거절될 확률(specificity)이 같아야 함을 의미한다.

기준의 엄격함을 완화하여 위음성(FN)에만 초점을 두고, 남녀간, 장애인・비장애인간 민감도(sensitivity)만 동일하면 만족되는 기준(즉, 위 수식이 y=+일 때만 성립하면 됨)을 기회균등(equal opportunity)이라 한다.[72]

- 충분성($Y \perp A | R$): 예측(prediction)을 기준으로 한 정확도(precision과 NPV)가 남녀간, 장애인・비장애인간 동일하도록 하는 것이다.[73]

$$\boldsymbol{P(Y = +|R = r, A = a) = P(Y = +|R = r, A = b),}$$

$$\boldsymbol{\forall a, b \in A, r \in \{+, -\}}$$

보험사기탐지의 예에서는, 보험사기로 분류된 남성의 청구와 여성의 청구는 실제로도 사기일 확률(precision)이 같아야 하며, 보험사기가 아닌 것으로 분류된 남성의 청구와 여성의 청구는 실제로도 사기가 아닐 확률(NPV)이 같아야 함을 의미한다.

기준의 엄격함을 완화하여 위양성(FP)에만 초점을

69) 독립성은 group fairness, statistical parity, demographic parity, 분리성은 equalized odds, disparate mistreatment, conditional procedure accuracy equality, 충분성은 calibration, conditional use accuracy equality라고도 불린다 (Verma et al., *supra* note 68 at 3-5).

70) Barocas et al., *supra* note 68 at 9-10.

71) Id at 12-14.

72) Verma et al., *supra* note 68 at 4.

73) Barocas et al., *supra* note 68 at 14-15.

두고, 남녀간, 장애인·비장애인간 정밀도(precision)만 동일하면 성립하는 기준(즉, 위 수식이 r=+일 때만 성립하면 됨)을 예측동등(predictive parity)이라 한다.[74]

이 세 공정성 지표를 동시에 충족하는 것은 불가능하므로,[75] 어떤 지표를 금융AI 모델에 의한 부당한 차별 여부의 기준으로 삼을지 고민이 필요하다. 우선 자질(merit)을 전혀 고려하지 않고 결과적 평등만을 추구하는 독립성 기준을 따르기는 어렵다. 독립성은 오히려 더 차별적인 결과를 가져올 수 있다. 억지로 비율만 맞추려고 한 쪽 성별에 대해서만 자질 없는 사람들에게 많이 대출할 경우, 결국 그 성별의 채무불이행률만 높아져 지속적 편향을 야기하고 차별을 오히려 심화할 수 있다.[76]

자질을 고려하며 결과의 동등보다는 정확도/오류율의 동등을 추구하는 기준(분리성, 충분성) 중에서는, (1) 계약심사 및 조건 산정(여신심사, 신용평가, 보험계약심사 등)에 있어서는 자질이 있는 사람 중 거절되는 비율(= 1 - sensitivity)이 한쪽 성별(예컨대 여성)이나 장애인에게서만 높아서는 아니 되는 것이 관심사이므로, 분리성, 보다 정확히는 기회균등이 가장 핵심적인 판단기준이 되어야 할 것이고, (2) 사기 또는 이상거래 탐지(특히 손해사정 등 보험금 지급심사)에 있어서는 사기로 판명된 사람 중 무고한 사람의 비율(= 1 - precision)이 한쪽 성별(예컨대 남성)이나 장애인에게서만 높아서는 아니 되는 것이 관심사이므로, 충분성, 보다 정확히는 예측동등이 가장 핵심적인 판단기준이 되어야 할 것이다.[77] 특히 분리성 기준은 두 그룹의 수용자반응특성 곡선(Receiver Operating Characteristic curve; "ROC curve")이 만나는 지점에서 충족되며 이미 모델 훈련이 완료된 이후에 사후처리(postprocessing) 방식으로 조정할 수 있으므로 개발자가 구현하기 쉽다는 뚜렷한 장점도 있다.[78]

3) 추가적 과제

이와 같이 사후적 재량에 맡겨져 있던 개념들을 상당 부분 수리적 지표로 변환함으로써 AI모델의 훈련단계에서부터 알고리듬 개입이 가능해지고, 객관적이고 불편부당한 사후 검증도 가능해지며, 사전적인 법적 예측가능성도 제고할 수 있다. 다만 현실에 적용하기 위해서는 다음 사항의 추가 검토가 필요하다.

- 현실에서는 정확성 지표를 다른 집단 상호간 완벽히 일치시키는 것이 불가능하므로, 양 집단 간의 정확성 지표의 차이가 일정 오차범위(slack; ε) 내에 들어오는지 여부를 기준으로 차별 여부를 판단하게 된다.[79] 따라서 이 허용범위를 어떤 기준(예컨대 양 성간 정확성 지표 차이가 ε 이하가 되도록 할지, 아니면 여성의 정확성 지표를 남성의 정확성 지표의 1-ε 이상이 되도록 할지) 및 범위(ε = 10% 또는 20% 등)로 정할지에 대한 규범적 결단이 필요하다.[80]
- 상기 공정성 지표는 분류(classification)의 문제, 즉 예측($\hat{y}$)이 이산(discrete)일 때(대출승인 대 거절, 부정거래 판정 등) 잘 적용된다. 회귀(regression)

74) Verma et al., *supra* note 68 at 3-4.

75) Barocas et al., *supra* note 68 at 18-21.

76) Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold & Rich Zemel, *Fairness Through Awareness, in* Proc. 3rd ITCS (2012), 214-226 (이러한 현상을 "self-fulfilling prophesy"라 칭함).

77) 이 아이디어는 서울대학교 산학협력단의 2021년 금융분야 인공지능 활성화를 위한 가이드라인 등 마련을 위한 금융위원회 연구용역 보고서에도 반영되어 정책 제안에 포함되었음을 밝힌다.

78) Barocas et al., *supra* note 68 at 13-14.

79) Barocas et al., *supra* note 68 at 18-21.

80) 예컨대 미국 공평고용기회위원회(EEOC), 노동부(DOL), 법무부(DOJ), 구 공무원위원회(CSC)의 Uniform Guidelines on Employee Selection Procedures는 독립성 기준을 취하고 어떤 인종, 성별, 민족 비율이 다른 인종, 성별, 민족 비율의 1-ε 미만이면 차별적 효과(adverse impact)의 증거로 간주하면서 ε = 20%로 정하고 있다(Four-Fifths Rule 또는 80% Rule) (Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger & Suresh Venkatasubramanian, *Certifying and Removing Disparate Impact*, ACM Proc. 21st SIGKDD (2015)).

의 문제, 즉 예측($\hat{y}$)이 연속(continuous)인 경우(신용평가, 보험료율의 산정 등)에 공정성을 어떻게 측정, 평가할지 여부에 대해서는 아직 일정한 합의에 이르지 못한 채 다양한 연구들이 진행 중이다. 예컨대 분리성에 상응하는 기준으로 유사한 자질(merit)을 가진 사람들이 유사한 제곱오차(squared error) 내에 분포하도록 하는 방안이 제안되기도 한다.[81] 다만 회귀의 경우에도 일정 구간별로 범주화하는 방식으로 분류와 동일한 접근이 가능하며, 이 경우 어떤 방식으로 범주화할지 기준에 대한 추가적인 논의가 필요할 것이다.

- 감독당국은 금융기관들이 이러한 공정성 지표들을 테스트할 수 있는 인프라를 개발, 구축하여 지원할 수 있을 것이다. 싱가포르 금융청(Monetary Authority of Singapore; "MAS")은 2021. 1. 6. 베리타스(Veritas), 즉 민관협력 금융AI 프로젝트의 첫 번째 결과물로 공정성 지표에 기한 공정성 평가 방법론과 오픈소스 코드를 공개했다.[82]

3. 금융AI에 대한 설명요구권, 정보제출권, 이의제기권

민간분야에서의 AI에 대한 투명성(transparency)의 요구는 상당히 빠르게 경성법화되고 있다. EU 일반정보보호법(General Data Protection Regulation (EU) 2016/679; "GDPR") 제22조는 (법적 효력이 있거나 유사한 중요한 영향이 있는) 자동화처리에만 기한 의사결정(a decision based solely on automated processing; 프로파일링 포함)의 경우 정보주체의 거부권, 개입요구권, 의견표시권, 이의제기권을 인정하고 있다. 이 조항은 2020. 8. 4. 시행된 신용정보법 제36조의2에 계수되어, 이에 따라 개인인 신용정보주체는 개인신용평가회사나 금융기관에 대해서 (1) 설명요구권(개인신용평가, 은행이나 금융투자업자의 신용공여, 신용카드・시설대여・할부금융거래 등 금융거래의 설정・유지 여부, 내용 결정, 이에 관한 계약의 청약・승낙 여부 결정에 자동화평가를 하는지 여부, 자동화평가를 하는 경우 결과, 주요 기준, 이용된 기초정보의 개요 설명 요구), (2) 정보제출권(결과 산출에 유리하다고 판단되는 정보 제출), (3) 이의제기권(기초정보의 정정・삭제, 재산출 요구)이 있다. 개인정보보호위원회가 2021. 1. 6. 입법예고한 개인정보 보호법 개정안 제37조의2는 개인정보처리자가 자동화된 개인정보 처리에만 의존하여 특정 정보주체에게 개별적으로 법적 효력 또는 생명・신체・정신・재산에 중대한 영향을 미치는 의사결정을 행한 경우 정보주체에 거부・이의제기・설명요구권을 부여한다. 미국 연방거래위원회(Fair Trade Commission; "FTC")는 2020년 자동화툴(automated tool)이나 알고리듬에 의한 의사결정(algorithmic decision-making)의 경우 공정신용보고법(Fair Credit Reporting Act, 15 U.S.C. §1681; "FCRA") 등 적용에 대한 해석을 제시했는데 그 핵심은 알고리듬에 기해 소비자에게 상품 제공을 거부할 때 이유를 설명하고 알고리듬에 기해 소비자의 위험도 평가 시 핵심요소를 중요도 순으로 공개하라는 것이다.[83]

---

81) Richard Berk, Hoda Heidari, Shahin Jabbari, Matthew Joseph, Michael Kearns, Jamie Morgenstern, Seth Neel & Aaron Roth, *A Convex Framework for Fair Regression*. arXiv:1706.02409 (2017).

82) MAS, *Veritas Initiative Addresses Implementation Challenges in the Responsible Use of Artificial Intelligence and Data Analytics* (2021), https://www.mas.gov.sg/news/media-releases/2021/veritas-initiative-addresses-implementation-challenges.

83) FTC, Using Artificial Intelligence and Algorithms (2020), https://www.ftc.gov/news-events/blogs/business-blog/2020/04/using-artificial-intelligence-algorithms. 그 밖에도 연방의회의 Algorithmic Accountability Act of 2019, 캘리포니아주의 Automated Decision Systems Accountability Act of 2021 (AB 13), 뉴저지주의 Algorithmic Accountability Act (AB 5430) 법안도 각각 발의된 바 있다. 뉴욕시의회는 2018년 Automated decision systems used by agencies라는 조례를 통과시켰으나 공공부문에 한하여 적용된다.

물론, 위 조항들은 현실에서 별로 적용되지 않을 것으로 보이므로 실제 큰 문제는 없을 수도 있다. 신용정보법의 경우 "자동화평가"는 "신용정보회사등의 종사자가 평가 업무에 관여하지 아니하고 컴퓨터 등 정보처리장치로만 개인신용정보 및 그 밖의 정보를 처리하여 개인인 신용정보주체를 평가하는 행위"라 정의되어 있고(제2조 제14호), GDPR이나 개인정보 보호법 개정안도 오로지 자동화된 처리에만 의할 경우에 적용되므로, 사람이 조금이라도 의사결정에 관여하면 적용이 배제된다. EU의 가이드라인은 사람의 관여를 작출(fabricating human involvement), 즉 사람의 관여가 실제 결과에 영향이 없이 자동적으로 생성된 프로파일을 반복적으로 적용할 경우 위 조항에서 빠져나올 수 없다 하나[84] 실제 평가 과정에서 사람을 일부라도 관여시키는 것은 어렵지 않을 것이다.

다만, 신용정보법 조항이 금융기관이 가용한 다양한 기술적 방식 중 "자동화평가"를 선택했다는 이유만으로 느닷없이 정보주체에 공공부문에 대한 알 권리에 준하는 권리를 부여하는 것은 실증적 근거에 기인한 것이 아니라 "블랙박스 알고리듬에 의한 자율성 침해"라는 실증도 반증도 될 수 없는 관념론에 기인하고 있다는 점이 문제이다. 이러한 프레임은 전통적 회귀분석 기법을 사용하면 자동화평가가 아니니까 투명할 필요가 없고, 회귀분석 시 분산을 통제하는 페널티(penalty)를 추가하여 라쏘(lasso; L_1 regularization)나 릿지(ridge; L_2 regularization)와 같은 정규화된(regularized) 회귀분석이 되면 기계학습의 일종이 되어 자동화평가니까 투명해야 한다는 식의 유치하고 기술중립성(technological neutrality)도 결여된 접근으로 빠질 수 있다. 자동화평가 시 인간의 개입(human-in-the-loop)을 확보하겠다는 본래의 취지와 달리 결국 쉬운 회피 과정에서 자동화평가라는 본래 활용 가능했던 기술수단의 활용만 억지하는 결과밖에 가져오지 못할 우려가 있다.

금융산업의 전통적 계량적 방법론들과 비교할 때 이러한 AI에 대한 투명성 규제가 대체 어떤 정책적 목표를 위해 존재하는지 원점에서부터 고민해 보면, 결국 투명성 규제는 낮은 신용평가를 받거나 대출을 거절당한 금융소비자가 실제 느끼지도 않을 "자율성 침해"가 아닌 실제 느낄 수 있는 공정성에 대한 의혹을 풀어주는 수단이 되어야할 것으로 생각된다. 구체적으로, 금융소비자가 금융소비자보호법 제15조가 금지하는 차별 여부를 검증하기 위한 수단이 되어야 할 것이다. 그렇다면 "자동화평가" 여부라는 기술적 방식을 기준으로 규제 적용 여부를 결정할 것이 아니라, (1) 성별 등 생래적이거나 장애 등 쉽게 바꿀 수 없는 특성값(features)에 의해 불이익을 받았는지 여부는 공정성 여부에 문제제기하기 위한 기초이므로 이들 특성값들이 분류에 미친 영향에 관한 설명요구권을 인정하되, (2) 소셜미디어에 남긴 글, 구매이력 등 쉽게 바꿀 수 있는 특성값들은 공정성과도 관련이 없고, 이들에 대한 설명요구권을 함부로 인정하면 가짜 소셜미디어를 만들어 좋은 신용평가를 받을 수 있게 해주는 컨설팅 서비스가 횡행하는 등 전략적 행동(gaming)으로 인해 분류모델이 무력화될 수 있으므로[85] 설명요구권을 부인하는 등, 특성값을 기준으로 규제 유무를 결정해야 하며, 금융소비자보호법 제15조의 차별금지사유 해당 여부가 중요한 판단기준이 될 것이다.

마지막으로, 설명의 대상이 되어야 하는 것은 알고리듬이 아닌 (데이터로 훈련된) 모델이라는 점을 강조하고자 한다. 사실 기계학습(특히 지도학습) 알고리듬 자체는 대체로 모델의 예측값과 데이터 중 실제 결과값 간의 손실함수(loss function)을 최소화[86]하는 방

84) Article 29 Data Protection Working Party, *Guidelines on Automated Individual Decision-Making and Profiling for the Purposes of Regulation 2016/679* (2018), 21.

85) Joshua A. Kroll, Joanna Huey, Solon Barocas, Edward W. Felten, Joel R. Reidenberg, David G. Robinson & Harlan Yu, *Accountable Algorithms*, 165 U Pa L Rev 633, 654 (2017).

86) 연속값을 예측하는 회귀모델의 경우 최소제곱추정(LSM)을 위해 오차제곱합(sum of squared errors; "SSE")을 최소화하고, 이산값을 예측하는 분류모델의 경우 최대가능성추정(MLE)을 위해

식으로 모델을 데이터에 피팅(fitting)하는 것이다. 이는 직관적으로 사람이 다른 사람을 오랫동안 관찰하며 자신이 생각하는 그 사람에 대한 인상을 그 사람의 실제 행동거지에 따라 맞춰가며 바꿔나가는 과정과 별로 다를 것이 없다. "블랙박스"가 아니며 특히 개발자는 완벽하게 이해할 수 있다. 개발자도 때때로 명확하게 이해하지 못할 수 있고 분류대상자에게도 설명이 필요할 수도 있는 것은 알고리즘이 아니라 모델(model)이다. 모델의 본질은 단순한 선형모델의 경우 각 데이터의 패턴을 반영하여 도출된 파라미터(parameter)(전통적 회귀분석에서의 계수(coefficient)와 절편(intercept))이고, 심층학습(deep learning)의 경우 각 노드(node)의 가중치(weight)와 편이(bias)를 의미한다. 즉, 각각의 특성값(feature)(예컨대 카드 발급 건수)이 1이 늘어날 때마다 예측값(예컨대 연체율)이 얼마만큼 바뀌는지를 벡터(vector)나 행렬(matrix) 형태로 구조화한 것이 모델이며, 이는 결국 원 데이터의 패턴 내지 각 특성값 간의 상관관계(correlation)를 반영한다. 쉽게 말해, 대출이 거절된 소비자가 듣고 싶은 것은 "당신은 랜덤포레스트 알고리듬에 기해 대출거절되었다"는 이야기가 아니라, "당신의 연체율이 특정 성별이라 몇 % 높게 예측되고, 장애인이라 몇 % 높게 예측되어 대출거절되었다"는 이야기이며, 이런 이야기를 들어야 차별에 대해서 문제제기할 수 있게 된다. 비선형모델들이 다 이런 식으로 설명이 가능하진 아니하여 설명가능한 상태로 변환하기 위한 다양한 설명가능한 AI(eXplainable AI; "XAI") 기법들이 제안되고 있다. 그러나 이러한 목적에 가장 충실한 기법은 예제기반 설명(example-based explanations), 특히 반사실적 설명(counterfactual explanation)(예측을 달라지게 하는 가장 최소한의 변화 산출)이라는 공감대가 생기고 있는 듯하다.[87]

III. AI에 의한 금융소비자 보호: 6대 판매규제 준수를 중심으로

1. 들어가며

소셜미디어 텍스트 마이닝에 기한 신용평가의 예(신용정보법 제15조 제2항 제2호 다목 참조)에서 보듯 AI는 포용금융(financial inclusion)을 위해 활용될 잠재력도 있으나, 이 장에서 살펴보고자 하는 것은 규제 준수 및 감독을 촉진하기 위한 정보기술 솔루션의 사용(regulatory technology; "RegTech")이다. 이 중 Basel III 등 건전성 규제 준수를 위한 은행의 기술 활용도 주요 쟁점이나,[88] 이 장에서는 소비자보호 규제 준수를 위한 기술 활용을 금융상품판매업자 등의 준수(compliance technology; "CompTech") 및 감독당국의 규제와 감독(supervisory technology;

교차엔트로피오차(cross entropy error; "CEE")를 최소화.

87) Sandra Wachter, Brent Mittelstadt & Chris Russell, *Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR*, 31 Harv J L & Tech 841 (2017).

88) 예컨대 Basel III에 따라 은행법 제34조, 은행업감독규정 제26조 제1항 제2호는 향후 30일간 순현금유출액에 대한 고유동성자산의 비율(유동성커버리지비율(liquidity coverage ratio; "통합LCR"))을 100% 이상(외은지점은 60% 이상; COVID-19로 85%로 한시 완화)으로(단기 유동성 규제), 외국환거래법 제11조 제2항, 시행령 제21조 제4호, 은행업감독규정 제63조의2는 외화자산・부채에 대하여 향후 30일간 순현금유출액에 대한 고유동성자산의 비율("외화LCR")을 80% 이상으로(COVID-19로 70%로 한시 완화)(단기 외화유동성 규제), 은행법 제34조, 은행업감독규정 제26조 제1항 제4호는 안정자금조달필요금액에 대한 안정자금가용금액의 비율(순안정자금조달비율(net stable funding ratio; "NSFR"))을 100% 이상으로(중장기 유동성 규제) 유지하도록 하고 있다. 통합LCR과 외화LCR의 경우 은행업감독업무시행세칙 [별표 3의6] 유동성커버리지비율 산출 기준에 따라 영업일별로 산정하여 매월 평잔을 감독원장에게 보고(미달(예상) 시 즉시 보고)해야 하고(제58조), NSFR의 경우 동 세칙 [별표 3의10] 순안정자금조달비율 산출 기준에 따라 최소 분기별로 산출하여 감독원장에게 보고(미달(예상) 시 즉시 보고)해야 하는데(제29조), 이 과정에서 RegTech을 활용하여 보고와 검토에 필요한 인력과 비용을 절감할 수 있다.

"SupTech")의 측면에서 전자에 좀더 비중을 두어 살펴본다.

RegTech라는 용어는 영국 금융영업감독청(Financial Conduct Authority; "FCA")이 2014년부터 시작한 "Project Innovate"라는 정책 프로젝트 과정에서 (규제샌드박스(regulatory sandbox)라는 용어와 함께) 만들어 낸 개념이다. FCA는 서브프라임 사태 이후 도입된 무거운 규제들로 인한 막대한 보고 및 감독 비용을 효율화하기 위해 RegTech 개발이 필요함을 역설하면서 시장의 의견수렴을 거쳐 2016년 RegTech의 핵심을 (1) 효율성과 협업(대안적 보고 방식, 공유 유틸리티, 클라우드, 온라인 플랫폼), (2) 집약・표준화・이해(시맨틱 기술과 데이터 포인트 모델, 공유 데이터 온톨로지, API, 로보 가이드라인), (3) 예측・학습・간소화(빅데이터, 리스크 및 컴플라이언스 모니터링, 모델링/시각화, 기계학습/인지기술), (4) 새로운 방향(블록체인/분산원장, 빌트인 컴플라이언스, 생체인증, 시스템 감시 및 시각화)으로 제시했다.[89] 이후 세계 곳곳에서 RegTech의 개발 및 적용에 대한 검토가 이루어졌는데, (건전성 규제에 따른 은행의 보고 및 감독기관들의 검토를 위한 RegTech의 활용 이외에는) 자금세탁방지와 테러자금조달금지(AML/CFT) 준수를 위한 RegTech의 활용이 상대적으로 활발한 듯하다.[90] 그러나 우리나라는 AML/CFT가 이미 특정금융정보법 및 테러자금조달금지법에 따른 금융정보분석원(FIU)의 시스템 내에서 구현되고 있어, 이보다는 금융소비자보호법상 판매규제 준수를 위한 AI 활용이 현안이라 할 수 있다. 이에 대해서는 호주 증권투자위원회(Australian Securities & Investment Commission; "ASIC")의 여러 시도가 특히 주목된다. ASIC은 2018-19, 2019-20 회계연도 중 RegTech 솔루션의 개발・적용을 위한 정부 예산지원을 받아, 다양한 RegTech 관련 시연 및 검토를 하였다.[91] 2018-19 회계연도 중에는 금융상품 권유, 투자자문, 음성 분석 및 텍스트변환, 인허가 관련 지침에 대한 챗봇 등 기술적 보조에 대해 검토하였고,[92] 2019-20 회계연도 중에는 금융상품 권유, 음성 분석 연구를 심화하면서, 데이터 자동화 및 프로세스 워크플로우, 자연어처리를 이용한 핵심 투자안내서 추출, 증거문서 평점 시스템 등을 검토하였다.[93]

다만, 우리나라의 경우 금융소비자보호법이 2021. 3. 25. 시행되자 준수를 위한 AI의 활용이 검토되기보다는 오히려 AI 등 기술의 적용이 중단되는 혼란이 초래되기도 했다. 동법상 설명의무를 이행하기 위한 녹취와 설명서 발급 프로세스가 미처 적용되지 않아 스마트텔레머신(STM)(차세대 ATM, 스마트 키오스크)과 비대면 상품(STM 통한 계좌개설, 전자적 투자조언장치 추천 집합투자 가입 등)이 중단되었다고 한다.[94] 다만 동법 시행령 제14조 제3항이 설명서 제공 방식으로 서면교부, 우편 뿐 아니라 이메일과 SMS도 인정하고 있으므로 이 문제는 기능 미적용으로 인한 일시적 혼란으로 보이고, 역시 향후 어떤 방식으로 AI를 활용하여 6대 판매규제를 준수할 수 있을지가 진정한 쟁점으로 자리 잡을 것으로 예상된다.

---

89) FCA, *Feedback Statement: Call for Input on Supporting the Development and Adopters of RegTech*, FS16/4 (2016), https://www.fca.org.uk/publication/feedback/fs-16-04.pdf.

90) 예컨대 미국 연방준비제도이사회(Federal Reserve Board of Governors), 싱가포르 금융청(MAS), 이탈리아 중앙은행(Banca d'Italia)은 자연어처리를 통해 다량의 의심거래보고(STR) 등 비정형데이터로부터 정보를 추출하고 이로부터 이상거래를 감지하는 시스템을 구축하였다고 한다(Financial Stability Board, *The Use of Supervisory and Regulatory Technology by Authorities and Regulated Institutions* 48-9, 57 (2020), https://www.fsb.org/wp-content/uploads/P091020.pdf).

91) ASIC, *ASIC's RegTech Initiatives 2018-19* (Report 653) 3 (2019), https://download.asic.gov.au/media/5424092/rep653-published-20-december-2019.pdf.

92) Id at 3-4.

93) ASIC, *ASIC's RegTech Initiatives 2019-20* (Report 685) 3 (2021), https://www.asic.gov.au/regulatory-resources/find-a-document/reports/rep-685-asic-s-regtech-initiatives-2019-20/.

94) 한국경제, "준비 덜 된 '금소법' 25일 시행 … 비대면 금융상품 판매 중단 '혼란'" (2021. 3. 25.).

2. 금융기관의 소비자 보호 규제 준수를 위한 AI 활용

금융소비자보호법이 시행되면서 개별 금융업법에서 일부 금융상품에 대해서만 적용하던 6대 판매규제를 모든 금융상품에 확대 적용하게 되었다. 6대 판매규제 중 (1) 적합성원칙(suitability)(제17조)은 소비자의 재산상황, 투자경험・목적 등을 고려하여 부적합한 상품을 권유하지 못하게 하는 것, (2) 적정성원칙(appropriateness)(제18조)은 소비자에게 권유하지 않은 상품이 소비자의 재산상황, 투자경험・목적 등에 비추어 부적정할 경우 그 사실을 고지, 확인하라는 것, (3) 설명의무(제19조)는 상품 권유 또는 소비자 요청 시 상품의 중요사항을 설명하라는 것, (4) 불공정영업행위금지(제20조)는 상품 판매 시 구매 강요, 담보・보증・편익 요구 등 소비자 권익을 침해하지 말라는 것, (5) 부당권유금지(제21조)는 상품 권유 시 소비자의 오인을 유발하지 말라는 것, (6) 광고규제(제22조)는 광고에 일정한 내용을 포함하도록 하거나 금지하는 것을 각 의미한다.

이들 6대 판매규제 준수의 효율화를 위해 AI는 핵심적인 역할을 수행할 것으로 기대된다. 6대 판매규제는 주로 소비자의 정보불균형(information asymmetry)을 시정하고 부수적으로 위험을 관리하기 위해 도입되었지만, 규제법령 전반에 흩어져 있는 다른 다양한 의무공개(mandatory disclosure) 규제들과 마찬가지로 실제 소비자들이 정작 나열식으로 제공되는 설명을 귀 기울여 듣지 않은 채 계약을 위해 지루하게 끝나기만을 기다려야 하는 요식절차로 전락할 수 있고, 오히려 이용자의 의사결정을 위해 꼭 필요한 핵심적이고 유용한 정보만 밀어낼(crowd out) 수 있으며, 거래비용(transaction cost)만 높여 금융소비자의 후생을 떨어뜨리는 효과를 초래할 수 있다.[95] 다량의 데이터 분석과 AI의 활용은 자동화(automation)를 통해 거래비용을 획기적으로 낮추는 한편, 소비자의 행태 및 개별 특성에 대한 실증적이고 깊이 있는 이해를 바탕으로 한 넛징(nudging)[96] 내지 법적용의 개별화(personalization)[97]를 가능하게 함으로써, 이러한 부작용을 완화할 수 있다. 구체적으로, AI는 다음과 같은 방식으로 활용될 수 있을 것이며, 각기 다른 법적 쟁점이 있다.

(1) 고객상담 녹취 정보를 텍스트변환(Speech-to-Text; "STT")한 후 자연어처리를 통해 분석함으로써 불완전판매 여부를 감지(detection)

KB국민은행이 금융소비자보호법 시행을 이틀 앞둔 2021. 3. 23. STT 기반 실시간 불완전판매 분석 시스템을 국내 최초로 구축했다고 발표한 이래,[98] 다른 여러 금융기관들도 활발히 시스템을 구축하고 있는 것으로 보인다. 기술적으로는 결국 음성을 STT로 텍스트 변환한 후, 텍스트분류(text classification), 즉 텍스트를 입력받아 정상판매, 불완전판매 중 어떤 클래스(class)에 속하는지, 불완전판매라면 설명 미비, 불공정영업행위, 부당권유행위, 부당 광고 등 세부 유형 중 어디에 해당하는지 분류하는 모델을 훈련시키는 형태가 될 것이다. 이러한 텍스트 분류모델은 스팸메일 필터링(spam filtering)과 같은 이상감지(anomaly detection), 감성분석(sentiment analysis), 의도분석(intent analysis) 등 다양한 용도로 이미 사용 중이며 나이브베이즈(Naive Bayes) 같은 단순한 알고리듬부터 순환신경망(RNN), 장단기메모리모델(LSTM) 같은 향상된 알고리듬까지 다양한 방법론이 제시되고 있으므로 (만약 충분한 데이터만 있다면) 기술적으로 구현하는 데에는 별 문제는 없다.

---

95) Omri Ben-Shahar & Carl E. Schneider, *The Failure of Mandated Disclosure*, 159 U Penn L Rev 647, 729-742 (2011).

96) Richard H. Thaler & Cass R. Sunstein, *Nudge: Improving Decisions about Health, Wealth, and Happiness* (2008).

97) Omri Ben-Shahar & Ariel Porat, *Personalizing Mandatory Rules in Contract Law*, 86(2) U Chi L Rev 255 (2019).

98) 조선비즈, "국민은행, AI 금융상담 시스템 구축 … 불완전판매 실시간 점검" (2021. 3. 23.).

그러나 몇 가지 법적 쟁점이 있다. 첫째, 고객이 대화 상대방인 대화를 대량으로 녹취해야 하는 문제가 있다.[99] 개인정보보호법이나 신용정보법에 따라 비식별화하면 물론 좋겠지만 개인정보보호법이나 신용정보법이 동의 면제를 위해 현재 요구하는 것은 비식별화를 위한 최선의 노력(best effort)이 아니라 완벽한 비식별화인데[100] 음성이나 텍스트 같은 비정형데이터(unstructured data)를 완벽히 비식별화하는 것은 만만한 일이 아니다. 특히 프로세스를 실시간으로 구현하려면 수작업을 배제하고 알고리듬적으로만 구현해야 하는데, 이 경우 완벽한 비식별화는 거의 불가능하다. 더욱이 통신비밀보호법의 해석과 관련하여 회사가 판매직원과 고객 간의 대화내용에 대해서 타인이 아닌 대화의 당사자에 해당하는가의 법적 쟁점이 완벽하게 해소된 것도 아니고, 통신비밀보호법 위반이 아니더라도 일방이 동의 없이 녹취하면 결국 판례상 민사불법행위에 해당하므로,[101] 어쨌든 동의는 받아야 한다.

결국 고객으로부터 개인(신용)정보 수집, 이용에 대한 동의 및 (적어도 민사불법행위에 해당하지 않기 위한) 녹취 동의를 겸하는 동의를 받아야 할 터인데, 가뜩이나 지루한 고객 실사와 설명을 앞두고 있는 금융소비자에게 실은 별 의미도 없는 긴 개인(신용)정보 고지를 듣게 하는 것이 올바른 대접인지 의문이다. 상품 권유의 경우 별도 고지를 통한 동의가 필요할 수 있을 뿐 아니라(개인정보보호법 제26조 제4항), 이용목적도 판매 뿐 아니라 텍스트 변환, 기계학습 모델에 대한 투입, 다른 6대 판매규제 준수를 위한 모델로의 응용, 타 금융상품판매업자등과의 데이터 결합 등 다양해서 이를 일일이 고지하는 것이 만만치 않을 것이다.

이런 것을 고지할 시간에 금융상품의 위험에 대해 조금이라도 더 설명하는 것이 낫다고 생각된다. 개인정보보호위원회, 금융위원회가 함께 녹취가 금융소비자보호법상 명시되어 있는 점(제17조 제2항, 제18조 제2항, 제19조 제2항)을 근거로 "법령상 의무를 준수하기 위하여 불가피한 경우"(개인정보보호법 제15조 제1항 제2호, 제17조 제1항 제2호, 신용정보법 제16조 제2항 제1호), "이 법 또는 다른 법률에 따라 제공(이용)하는 경우"(신용정보법 제32조 제6항 제10호, 제33조 제1항 제4호)로 보아 개인(신용)정보 수집・이용・제공 동의를 면제하는 취지의 가이드라인을 발하고, 근본적으로는 관련 법령을 정비해서 녹취에 대한 간략한 고지만으로 동의를 얻을 수 있게 할 필요가 있다.

둘째, 분류모델을 제대로 훈련・검증・시험하려면 정상판매와 불완전판매가 레이블링(labeling)되어 있는 다량의 텍스트 데이터가 필요하다. 이상적으로는 정상판매, 불완전판매 중 판단을 내린 법원 판결이나 감독당국의 해석례, 실제 분쟁 사례 및 그 결과 등의 데이터, 즉 공신력 있는 레이블링이 있는 데이터가 충분히 존재하면 좋겠지만 현실은 그렇지 않다. 결국에는 금융상품판매업자등이나 솔루션 개발업체가 레이블링 되어 있지 않은 텍스트 데이터에 각자의 직관적 판단에 따라 임의로 레이블링을 하여 이를 바탕으로 모델을 훈련할 수밖에 없을 것이다. 이 경우 AI 모델이 정상판매로 분류했다는 이유만으로 임직원에 대한 관리책임(금융소비자보호법 제16조)을 다하였다고 인정받거나 불완전판매에 대한 책임을 면할 수 있다고 하긴 쉽지 않을 것이다.

더욱이, 금융상품판매업자등 입장에서는 불완전판매가 아닌데 불완전판매로 분류하는 위양성(FP)이 발

99) 물론 판매직원의 발화도 대량으로 녹취해야 하는 문제가 있으나, 일단 전 판매직원으로부터 녹취 및 개인정보(음성) 수집, 이용에 대한 동의를 받는다는 것을 전제로 이에 대한 논의는 생략한다.

100) 훗날 비식별화를 위한 최선의 노력 시 2020년 개정으로 추가된 "개인정보의 추가적인 이용 또는 제공"(개인정보보호법 제15조 제3항, 제17조 제4항, 신용정보법 제32조 제6항 제9의4호)에 해당한다는 해석이 내려지면 달라지겠지만, 아직 이러한 해석이 내려진 바 없다. 비정형데이터가 개인정보보호법 제2조 제4호의 "개인정보파일"에 해당하지 않으므로 동법의 적용을 받지 말아야 한다는 주장이 있을 수 있으나(서울중앙지방법원 2016. 12. 15. 선고 2016고합538, 558(병합) 판결 참조) 좀 더 검토가 필요하다.

101) 수원지방법원 2013. 8. 22. 선고 2013나8981 판결, 서울중앙지방법원 2018. 7. 6 선고 2017가합548478 판결 등.

생하여 재차 설명 등을 하는 과정에서 소비자의 불만으로 상품판매가 어려워져 매출이 줄어드는 효과는 분명치 않을 수 있는 반면, 불완전판매인데 정상판매로 분류하는 위음성(FN)이 발생하여 법적 책임을 지게 될 위험은 가시적이라, 특이도(specificity)보다는 민감도(sensitivity) 위주로 모델을 피팅하게 될 수 있다. 그 결과 위양성이 다량 발생할 경우, 감사와 책임에 예민한 금융기관 임직원들이 AI 모델이 인간의 직관에 부합하지 않게 위양성을 발생시킨 경우라 할지라도, 상식에 따라 정상판매로 분류하는 것을 회피하고 그대로 양성(불완전판매)으로 처리할 가능성이 높다. 또한 금융기관 스스로도 "AI가 불완전판매로 분류했는데 사람이 묵살했다"는 정황이 불완전판매 소송 등에서 불리하게 작용할 수 있어 AI 모델의 위양성 분류를 맹목적으로 따를 가능성이 높다. 결국 사람이 분류하는 것보다 높은 빈도로 불완전판매로 분류되고 이미 충분히 제대로 설명을 들은 소비자에게 계속 다시 설명하여 보완하는 과정에서 도리어 불편을 끼칠 수 있을 것으로 전망된다.

위 문제를 해결하기 위해서는, 각 금융상품판매업자등이 각자 모델을 마련하는 형태는 한계가 있고, 금융업별 협회 등의 주도로 각 텍스트 별 금융소비자들의 충분한 이해 여부를 묻는 설문을 병행하는 등의 실증적인 방식으로 레이블링을 하여 이를 토대로 금융상품판매업자등이 전체적으로 활용할 수 있는 사전훈련된 모델(pretrained model)을 마련할 필요가 있다. 각 금융상품판매업자등이 이를 모듈 삼아 전이학습(transfer learning)을 통해 자신의 상품의 특수성을 반영하는 파인튜닝(fine-tuning)을 함으로써 맞춤형 분류모델을 쉽게 만들 수 있을 것이다. 협회 등은 데이터허브를 통해 각 금융상품판매업자등에 서서히 축적될 제대로 레이블링된 데이터(즉, 판결, 분쟁조정, 검사 결과 등과 매칭되어 있는 데이터)를 지속적으로 모아(앞서 살펴본 개인정보법 등 개정으로 동의가 면제되면 이 프로세스에 큰 도움이 될 것이다), 이를 토대로 분류모델을 더욱 개선시킬 수 있을 것이다.

(2) 소비자의 재산상황・구매목적・경험 등을 확인하여 적합성・적정성 여부를 분류하는 고객실사(Know Your Customer; "KYC") 프로세스를 자동화

금융소비자보호법은 "면담・질문 등을 통하여" 적합성・적정성 여부의 판단을 위한 정보를 파악하라고 하고 있는데(제17조 제2항, 제18조 제1항), 컴퓨터나 태블릿 화면 클릭 등을 포함한 다양한 확인 수단이 여기에 포함될 수 있다고 해석된다면 이러한 고객실사 프로세스도 자동화될 여지가 클 것이다. 다만, 이 프로세스도 앞서 살펴본 불완전판매 감지와 같이 텍스트분류(text classification)의 일종이 될 것이고, 마찬가지로 레이블링된 데이터 부족, 편향 등의 문제로 개별 금융상품판매업자등보다는 금융업별 협회 등의 주도로 통합모델을 마련해 나갈 필요가 있다.

다만, 적합・적정한 소비자를 부적합・부적정하다고 오분류하여 상품 권유를 포기하거나 부적정성을 고지하는 위양성(FP)의 문제는 불완전판매 감지에 비해 고객실사의 경우 더 심각할 수도 있고 그렇지 않을 수도 있다. 더 심각한 측면은 고객실사 과정에서의 위양성이 특정 소비자의 금융접근권을 부당히 약화시킬 수 있고, 이로 인해 공정성 문제도 야기할 수 있다는 점이다. 덜 심각한 측면은 위양성 분류가 매출 감소로 직결되기 때문에 각 금융상품판매업자들이 결코 특이도(specificity)를 무시할 수 없다는 점이다. 다만 어떠한 효과가 더 큰지 여부와 관계없이, 결국 소비자의 후생을 극대화하지 않는 방식으로 분류가 이루어질 유인이 발생한다는 점에서, 협회 등의 제3자의 통합모델 구축 방식이 여전히 유효한 대안일 수 있다.

아울러, 소비자들이 재산상황・구매목적・경험 등을 개별 상품 계약을 체결할 때마다 일일이 실사 당하는 것은 불편이 너무 크다. 감독 당국은 "과거 거래를 했던 소비자가 신규 거래를 하려는 경우에 과거에 소비자로부터 제공받은 정보와 적합성 판단기준에 변경이 없다면" "소비자 정보의 변경여부를 확인하는 절차로 적합성 평가를 갈음할 수 있다"는 가이드라인을

제시하였으나,[102] 이러한 편리함은 다른 금융상품판매업자등과 거래를 할 때에도 구현될 필요가 있다. 즉, 특정 금융상품판매업자등으로부터 적합성・적정성 평가를 받은 내역이 있으면 다른 금융상품판매업자등에 내역 전송을 요구한 후 (각 개별 상품별로 추가로 확인해야 하는 사항만 확인하는 방식의) 확인 간소화를 요구하는 권리를 인정해야 할 것이다. 이를 정보이동권(right to data portability)의 대상으로 구성하여 본인신용정보관리업자(마이데이터)를 통해 구현할 수 있도록 하는 것도 하나의 방법이 될 수 있다.

(3) 텍스트의 음성 변환(Text-to-Speech; "TTS")을 통한 설명의무 이행

이는 기술적으로 구현이 쉽고, 설명서를 서면이나 이메일, SMS 등으로 제공한 후(금융소비자보호법 제19조 제2항, 시행령 제14조 제3항 각호) 소비자가 특정 사항에 대한 설명만을 원하는지 확인하고, 그 내용을 TTS로 설명하는 것이 제19조 제1항의 "일반금융소비자가 이해할 수 있도록 설명"하는 것의 요건에 부합할 수 있다고 생각되므로,[103] 특별한 법적 문제도 없을 것이다.

(4) 챗봇(chatbot) 등 대화형 에이전트(conversational agent)의 활용

챗봇 등 대화형 에이전트의 최대 장점은 대화 과정에서 소비자별 특성을 파악하고 이를 반영하여 개별화된 맞춤 설명을 하도록 훈련할 수 있어 전반적으로 6대 판매규제 준수의 효율성을 향상시키고 규제로 인한 거래비용의 증가를 완화할 수 있다는 점일 것이다. 이를 활성화하기 위해 감독당국 등의 주도로 금융이해도와 관련한 다양한 분포를 가진 사람들이 대화형 에이전트를 통해 고객실사를 받고 설명을 받은 후 금융상품에 대한 제대로 된 이해를 얻었는지 여부를 실증적으로 측정하는 테스트베드(test best)를 마련하여, 이를 통해 대화형 에이전트를 검증하고, 이 검증을 통과한 대화형 에이전트를 사용해서 적합성・적정성 여부를 분류하는 고객실사부터 시작해서 설명에 이르기까지 판매규제를 준수할 수 있다고 해석할 수 있을 것이다. 실증의 방식은 설문 형태일 수도 있고, 크라우드소싱(crowdsourcing) 형태일 수도 있을 것이다. 검증에 사용된 실증자료는 불완전판매 소송에서 유력한 증거로도 활용될 수 있어 법적 불안도 상당히 완화할 수 있을 것으로 기대된다.

현재 대다수의 금융기관들의 대화형 에이전트는 사람처럼 자연스럽게 대화하는 오픈도메인 챗봇(open-domain chatbot)보다는 특정 주제에 대해서만 답변해 주는 목적지향형 챗봇(goal-oriented chatbot)의 수준에 머물러 있다. 그러나 결국 금융소비자들의 수요에 응하여 오픈도메인 챗봇으로(나아가 휴머노이드(humanoid)의 수준까지) 진화하려면 다량의 한국어 구어 말뭉치(corpus)로 모델을 훈련해야 한다. 말뭉치를 구성하는 각 대화내용에 개인정보가 포함될 경우 정보주체의 동의를 얻지 못했다면 이 대화내용을 가지고 챗봇을 훈련시키기 위해 비식별화를 해야 한다. 개인정보보호위원회는 2021. 4. 28. 오픈도메인 챗봇인 "이루다"가 타 앱을 통해 수집한 카카오톡 등 대화문장으로 훈련되어 이 중 한 답변을 선택하여 발화한 것이 수집 목적 외의 이용에 해당한다고 보아 (다른 위반사실에 대한 제재와 합쳐) 총 1억 330만원의 과징금과 과태료를 부과하고 시정조치를 명하였다. 오픈도메인 챗봇의 경우 대화내용을 가지고 단순히 자연어이해 모델을 훈련만 하는 것이 아니라 답변 데이터베이스에 포함시켜 챗봇 이용자에게 출력함으로써 사실상 불특정다수에 공개하는 결과로 이어진다는 점에서 비식별화의 방식(일단 공개되면 과학적

102) 금융위원회/금융감독원, "보도참고자료: 금융소비자보호법 시행 후 원활한 금융상품거래를 위해 판매자소비자가 알아야 할 중요사항을 알려드립니다"(2021. 3. 29.), https://www.fsc.go.kr/no010101/75631.

103) 위 보도참고자료는 "반드시 설명서를 구두로 읽어야 할 필요는 없으며 동영상 등 다양한 매체를 활용할 수 있습니다"라고 밝히고 있다.

연구 등 목적으로 한정할 수 없으니 익명처리를 해야 할지, 아니면 가명처리만으로 족한지 등)에 대해서 지속적인 법적 논란이 이어질 수밖에 없다. 비정형데이터의 특성상 대화문장의 수가 늘어나거나 지속학습(continual learning)이 이뤄지면 완벽한 비식별화가 더 어려워질 것이다. 이러한 법적 리스크로 인해 대화형 에이전트의 개발이 위축될 우려가 있으므로, 금융업별 협회 등의 틀 내에서 통합적인 금융업 말뭉치 구축(비식별화 포함) 및 제공 지원 사업을 통해 대화형 에이전트의 개발을 활성화할 필요가 있고, 보다 근본적으로는 비정형데이터에 대해서는 동의 없는 "개인정보의 추가적인 이용 또는 제공"(개인정보보호법 제15조 제3항, 제17조 제4항, 신용정보법 제32조 제6항 제9의4호)의 적용 범위를 넓혀 비식별화를 위한 최선의 노력(best effort)이 있었다면 사소한 미비가 있는 경우에도 제재 대상으로부터 제외하는 완화된 해석이 요망된다.

3. 감독당국의 소비자 보호 감독을 위한 AI 활용

한발 더 나아가 감독당국이 실시간으로 각 금융판매업자등의 고객 상담 녹취 파일 등을 취합하여 불완전판매 여부를 감지하는 SupTech 시스템을 구축해야 한다는 논의가 있을 수 있다. 그러나 앞서 살펴본 위양성(FP) 문제는 이 경우 더욱 심각해질 것이다. AI 모델이 인간의 직관에 부합하지 않는 위양성을 발생시킨 경우에, 감독당국의 담당자들이 감사 등 부담을 무릅쓰고 재량에 따라 이를 정상거래로 분류하여 무혐의로 처리하는 적극행정을 기대하기는 어렵다. 이러한 집중형 시스템은 금융소비자들이나 금융기관 임직원들의 프라이버시도 지나치게 침해하는 문제가 있으며 보안에도 취약할 수 있다. AML/CFT 수준의 중앙집중형 상시 감시체제를 운영하기보다는, 기반과 테스트 베드를 조성, 지원하는 역할이 바람직할 것이다. 구체적으로 앞서 살펴본 대로 (1) 협회 등의 불완전판매 감지 모델 또는 적합성・적정성 분류 모델 훈련・검증・시험 및 제공 지원, (2) 대화형 에이전트를 테스트할 수 있는 테스트 베드 구축 및 제공, (3) 대화형 에이전트의 고도화를 위한 금융업 말뭉치 구축 및 제공 지원 등을 통해 각 금융판매업자들의 CompTech 시스템 구축을 지원하고 촉진할 수 있을 것이다.

IV. 결어

금융AI로부터의 소비자 보호와 관련해서는, 금융소비자보호법 및 기타 관련 법령상 차별금지사유 및 부당성 기준을 모호한 사후적 판단에만 맡길 것이 아니라, 각 요건의 사전적 명징화를 통해 AI모델의 훈련・검증・시험 단계에서부터 비차별성을 구현할 수 있는 방법론을 공정 기계학습에서 논의되는 공정성 지표를 참고하여 제시해 보았다. 설명요구권은 현 신용정보법 조항처럼 자동화 의존 여부를 기준으로 적용할 것이 아니라, 공정성과의 관계가 약하고 전략적 행동의 대상이 될 수 있는 후천적, 가변적 특성값이 아닌 생래적, 불변적 특성값에 대해 설명을 요구하는 권리로 재편하는 방안을 제안하였다. 금융AI에 의한 소비자 보호와 관련해서는, 금융소비자보호법상 6대 판매규제의 준수를 AI모델을 통해 구현하는 방법과 법적 쟁점, 이를 정책적으로 촉진하는 방안에 대하여 논의하였다.

이를 통해, 결국 이미 전통적으로 계량적 방법론이 널리 쓰였던 금융시장에서 AI모델을 완전히 별개의 현상으로, 그것도 공공영역에 준하여 규제하자는 논의의 문제점을 드러내면서, 결국 AI라는 현상에 대응하여 법이 나아가야 할 가장 중요한 방향은 리스크의 과장과 규제의 신설・강화가 아닌 법체계 자체의 시대에 걸맞은 명징화와 고도화에 있음을 예시하고자 하였다.