ISSN 2288-4866 (Print)
ISSN 2288-4882 (Online)
http://www.jiisonline.org
J Intell Inform Syst 2022 September: 28(3): 259~278 http://dx.doi.org/10.13088/jiis.2022.28.3.259

# 인공지능 기반 금융서비스의 공정성 확보를 위한 체크리스트 제안: 인공지능 기반 개인신용평가를 중심으로

**김하영**
국민대학교 TED
스마트경험디자인학과
*(hayeongkim@kookmin.ac.kr)*

**허정윤**
국민대학교 TED
스마트경험디자인학과
*(yunlheo@kookmin.ac.kr)*

**권호창**
성균관대학교
트랜스미디어 연구소
*(kwonhc000@skku.edu)*

인공지능(AI)의 확산과 함께 금융 분야에서도 상품추천, 고객 응대 자동화, 이상거래탐지, 신용 심사 등 다양한 인공지능 기반 서비스가 확대되고 있다. 하지만 데이터에 기반한 기계학습의 특성상 신뢰성과 관련된 문제 발생과 예상하지 못한 사회적 논란도 함께 발생하고 있다. 인공지능의 효용은 극대화하고 위험과 부작용은 최소화할 수 있는 신뢰할 수 있는 인공지능에 대한 필요성은 점점 더 커지고 있다.

이러한 배경에서 본 연구는 소비자의 금융 생활에 직접 영향을 끼치는 인공지능 기반 개인신용평가의 공정성 확보를 위한 체크리스트 제안을 통해 인공지능 기반 금융서비스에 대한 신뢰 향상에 기여하고자 하였다. 인공지능 신뢰성의 주요 핵심 요소인 투명성, 안전성, 책무성, 공정성 중 포용 금융의 관점에서 자동화된 알고리즘의 혜택을 사회적 차별 없이 모두가 누릴 수 있도록 공정성을 연구 대상으로 선정하였다.

문헌 연구를 통해 공정성이 영향을 끼치는 서비스 운용의 전 과정을 데이터, 알고리즘, 사용자의 세 개의 영역으로 구분하고, 12가지 하위 점검 항목과 항목별 세부 권고안으로 체크리스트를 구성하였다. 구성한 체크리스트는 이해관계자(금융 분야 종사자, 인공지능 분야 종사자, 일반 사용자)별 계층적 분석과정(AHP)을 통해 점검 항목에 대한 상대적 중요도 및 우선순위를 도출하였다. 이해관계자별 중요도에 따라 세 개의 그룹으로 분류하여 분석한 결과 학습데이터와 비금융정보 활용에 대한 타당성 검증 및 신규 유입 데이터 모니터링의 필요성 등 실용적 측면에서 구체적인 점검 사항을 파악하였고, 금융 소비자인 일반 사용자의 경우 결과에 대한 해석 오류 및 편향성 확인에 대한 중요도를 높게 평가한다는 것을 확인할 수 있었다.

본 연구의 결과가 더 공정한 인공지능 기반 금융서비스의 구축과 운영에 기여할 수 있기를 기대한다.

**주제어** : 인공지능, 신뢰할 수 있는 인공지능, 인공지능 기반 금융서비스, 인공지능 기반 개인신용평가, 계층적 분석과정(AHP)

논문접수일 : 2022년 8월 26일 논문수정일 : 2022년 9월 24일 게재확정일 : 2022년 9월 26일
원고유형 : Regular Track 교신저자 : 허정윤

## 1. 서론

인공지능은 전통적인 제조업뿐만 아니라 금융, 의료, 교육, 서비스 등의 다양한 산업 분야에 도입되어 큰 혁신을 불러일으키고 있다(이정선 등, 2021). 그중에서도 금융 산업은 고빈도(high-frequency) 및 고품질(high-quality)의 정제된 데이터가 많고, 데이터 축적 속도 또한 상당히 빨라 서비스에 인공지능을 적극적으로 도입하고 있다(김지웅 등, 2013). 금융위원회는 금융업무의 특성을 반영하여 인공지능 도입 사례를 <표 1>과 같이 세 가지로 구분하고 있다(금융위

원회, 2021).

〈표 1〉 금융 분야 인공지능 도입 사례

| | |
|---|---|
| Front Office | 고객 상담 자동화로 고객 편의성 증대 |
| | 카카오뱅크 – 챗봇<br>NH 농협은행 – AI 은행원<br>신한금융투자 – 로보어드바이저 |
| Middle Office | 인공지능을 활용한 서비스 고도화 및 비즈니스 모델 혁신 |
| | 토스뱅크 – 개인신용평가모형<br>기업은행 – 전화금융사기 차단 앱<br>하나은행 – 인공지능 기반 자산관리 |
| Back Office | 내부적 업무 효율성 향상 및 비용 절감 |
| | 국민은행 – RPA(Robotic Process Automation) |

그러나 인공지능 기술에 대한 사회적 신뢰가 부족한 상황에서 차별과 편향 등의 문제가 발생하고 있다(Lepri et al., 2018). 애플과 골드만삭스가 출시한 '애플 신용카드(Apple Card)'가 같은 조건을 가진 남성의 신용카드 한도를 여성 대비 10배 이상 높게 적용한 사례가 대표적이다(KDB 미래전략연구소, 2021; Vigdor, 2019). 이처럼 인공지능에 대한 사회적 신뢰가 떨어지는 문제로 신뢰할 수 있는 인공지능 기반 금융서비스의 필요성이 커지고 있다(KDB 미래전략연구소, 2021).

이러한 배경하에 인공지능 기반 금융서비스의 신뢰를 제고하기 위한 실용적 연구로서 '인공지능 기반 개인신용평가'(Credit Scoring System, CSS, 이하 AI 기반 개인신용평가)의 공정성 개선방안을 모색하고자 하였다. 개인신용평가는 개인신용평가모형을 통해 금융 소비자의 신용정보를 통계적으로 분석하고 신용 활동과 관련된 신뢰도를 신용평점 형태로 산출하는 평가 방법이다(최성민, 2020). 개인신용평가의 결과는 카드 개설, 대출 승인 및 한도 결정 시에 기준이 되는 자료로 활용되어 소비자의 금융 생활에 직접적인 영향을 미치고 사회적 파급력이 크기 때문에 연구 대상으로 선정하였다. 또한, 소비자 보호 측면에서 개인신용평가에 가장 직접적인 영향을 주는 공정성 개선을 목표로 하여 인공지능 기반 금융서비스의 신뢰를 높이고자 하였다.

본 연구는 AI 기반 개인신용평가의 공정성 개선방안으로 서비스 운용의 전 과정을 데이터, 알고리즘, 사용자로 구분하였으며 문헌 연구를 통해 영역별 체크리스트를 제안하였다. 제안한 체크리스트는 점검 항목별 상대적 중요도를 도출하여 관련 이해관계자들의 인식 차이를 확인하기 위해 계층적 분석과정(Analytic Hierarchy Process, 이하 AHP)을 활용하였다. 이해관계자는 금융 분야 종사자, 인공지능 분야 종사자, 일반 사용자인 세 그룹으로 구분함을 통해 AI 기반 개인신용평가에 대한 다양한 관점을 수렴하고 사용자 친화적인 서비스 구축에 기여하고자 하였다. 마지막으로 연구 결과는 현재 운용되고 있는 개인신용평가 서비스와 관련해 가지는 시사점을 실용적 관점에서 논의하였다.

## 2. 관련 연구

개인정보 유출, 알고리즘의 편향과 차별이나 오남용 등 인공지능 시스템(이하 AI 시스템)의 문제들이 발생함에 따라 신뢰할 수 있는 인공지능에 대한 필요성이 커지고 있다(과학기술정보통신부, 2021). 인공지능의 신뢰성 구축을 위한 핵심 요소에는 <표 2>와 같이 안전성(Safety), 책무성(Accountability), 투명성(Transparency), 공정성(Fairness) 등이 포함되며 인공지능 윤리 가이

드라인이나 개발 지침 또는 연구자별로 주목하는 관점에 차이가 있고, 지속해서 보완되고 있다(과학기술정보통신부, 2021; 소순주, 안성진, 2021; 양희태 등, 2018; 장용석 등, 2020).

〈표 2〉 인공지능 신뢰성 핵심 요소 정의

| 핵심 요소 | 개념 설명 |
|---|---|
| 안전성 Safety | 인공지능의 판단 및 예측 결과로 인한 시스템 오작동이나 부작용 등이 사용자 환경에 악영향을 미치지 않도록 안전하게 서비스를 유지와 관리 할 수 있는 원칙 |
| 책무성 Accountability | 인공지능 기술과 관련된 정보 공유나 사고에 대한 책임을 명확히 정의하고, 보상 정도를 정확히 측정해 문제의 원인을 객관적이고 구체적으로 파악해 그 결과를 누구에게 물어야 하는지 등을 파악하는 책임의 원칙 |
| 투명성 Transparency | 인공지능에 의해 발생하거나 발생하지 않은 결과에 대한 이유를 명확하게 확인할 수 있는 원칙 |
| 공정성 Fairness | 인공지능의 데이터 처리 과정부터 활용되는 과정까지 특정 개인이나 집단에 대한 차별이나 편견이 없도록 공정하게 보장되어야 하는 원칙 |

네 가지 핵심 요소는 인공지능 기반 서비스의 운용 과정 중에 고려되는 시기와 목적에 차이가 있다. 안전성과 책무성의 경우 인공지능 활용으로 인해 발생한 문제 상황을 대처 및 해결하기 위한 목적에 집중되어 있다. 투명성은 서비스 배포 이후에 사용성이나 신뢰 제고를 위해 고려해야 하는 항목으로 안전성, 책무성, 공정성보다는 잠재 위험성이 낮다. 공정성은 인공지능 기반 서비스의 기획부터 배포 및 운영 단계까지의 서비스의 전 과정에서 중요하게 작용하며 상황에 따라 큰 사회적 파급력을 불러일으킬 수 있다. 따라서 본 연구는 인공지능 신뢰성 구성요소 중에서도 AI 시스템의 공정성에 집중하고자 하였다.

AI 시스템은 인간과 달리 기본 수준의 도덕적 인식을 의사결정에 스스로 반영할 수 없다(McCalman et al., 2022). AI 시스템이 윤리적 기준을 충족할 수 있게 만들기 위해서는 시스템 설계 과정에서 반영하고자 하는 윤리적 기준을 명확하게 정의하고 이를 시스템에 수학적으로 인코딩하는 과정이 필요하다(Microsoft, 2020). 이러한 배경으로 AI 시스템의 공정성을 개선하기 위한 기술적 접근 방식의 연구들이 꾸준히 늘고 있다. 관련 연구는 공정성을 수치로 측정할 수 있는 모델을 개발하는 연구가 대부분이며, 공정성 측정은 AI 시스템의 데이터와 예측 결과를 활용해 산출하는 방식으로 이뤄진다(Li & Chignell, 2022). 예를 들어, 대출 한도를 결정하는 AI 의사결정 시스템에서 공정성을 '성별에 상관없이 동일한 기준으로 예상 한도를 제공하는 것'으로 정의할 경우, 남성과 여성의 한도 예측 오류율을 비교하고 이를 공정성 개선을 위한 기준 지표로 활용할 수 있다.

기술적 관점에서 개인신용평가모형의 공정성을 개선한 문헌은 다음의 두 가지가 대표적이다. Fuster et al. (2017)은 미국의 전통적 신용평가 방식에 인공지능이 도입되면서 나타난 사용자 인종에 따른 금리 변화를 조사하였다. 전통적 신용평가 방식을 활용한 주택담보대출 사례에서 흑인과 히스패닉 사용자가 백인이나 아시안 사용자보다 불이익을 받은 경향이 나타났으며 인공지능의 도입으로 신용평가 결과의 공정성을 개선할 수 있다고 밝혔다(Fuster et al., 2017). "Kozodoi et al. (2022)는" 관련 연구가 제안한 AI 시스템의 목적에 따른 공정성 기준(fairness criterion)을 신용평가모형에 대입해 다양한 공정성 기준을 정의하고, 공정성과 이익 사이의 상충관계(trade-off relationship)가 어떻게 발전하는지

에 대한 실용적 결과를 제공했다(Hardt et al., 2016; Kozodoi et al., 2022) 이처럼 개인신용평가 모형의 공정성에 대한 선행연구는 변수를 조절해 알고리즘의 성능을 향상하는 기술적 접근 방식에 집중되어 있다. AI 시스템이 적용되는 분야나 상황에 따라 상이한 공정성 정의를 적용하는 접근 방식은 공정성을 평면적으로만 해석하게 하고 궁극적으로 AI 시스템의 공정성을 개선하기 위한 실용적 방안으로 활용하기에는 한계가 있다.

인간의 도덕적 기준을 수학적으로 인코딩하는 연구와 더불어 신뢰할 수 있는 인공지능 기반 금융서비스에 대한 원칙이나 지침을 개발하는 사회적 노력도 있다. 이러한 접근 방식은 알고리즘에 집중하는 것이 아니라 AI 시스템 설계자나 운영관리자가 인공지능 위험성을 인식하고 개선하는 데 도움을 주는 것을 목표로 한다(McCalman et al., 2022). 예를 들어, 싱가포르 통화청(Monetary Authority of Singapore, MAS)은 금융분야의 책임 있는 인공지능 활용을 위해 ‘FEAT (Fairness, Ethics, Accountability, Transparency) 원칙’을 발표했으며, 이를 고객 마케팅과 신용평가 서비스에 적용한 실증 사례 연구를 발표했다(MAS, 2020). AI 시스템 설계자와 운영관리자에게 일련의 질문과 지침이 포함된 가이드라인을 제공함으로써 AI 시스템의 공정성을 자체적으로 평가하고 모니터링할 수 있도록 돕는다. 국내의 경우 2021년에 금융위원회가 ‘금융분야 AI 가이드라인’을 발표했으며 기획 및 설계, 개발, 평가 및 검증, 도입, 운영 및 모니터링 등의 서비스 전 과정에서 고객 신뢰를 제고하기 위한 가이드를 제공하고 있다(금융위원회, 2021). 하지만 아직 서비스별로 차이가 있는 AI 활용사례 및 잠재 위험을 고려한 구체적인 서비스별 가이드라인은 마련되지 않은 상황이다. 인공지능 기반 금융서비스의 안전한 활성화를 위해 전체 서비스를 아우르는 넓은 범위의 가이드라인도 필요하지만, 분야와 서비스의 특성에 맞춰 현업에서 적용할 수 있는 구체적인 가이드라인도 필요하다.

이러한 맥락에서 서비스 확장성을 고려하여 AI 기반 개인신용평가 서비스의 기획 및 설계부터 사용자에게 적용되고 운영하는 전 범위에 걸쳐 공정성을 체계적으로 개선하는 방안을 설계하였다.

## 3. AI 기반 개인신용평가의 공정성 개선을 위한 방안 설계

### 3.1 국내 금융기관의 AI 기반 개인신용평가

국내 개인신용평가는 2021년 1월 1일을 기점으로 신용 등급제에서 신용점수제로 전면 전환되었으며, 개인신용평가회사(Credit Bureau, CB)가 1점부터 1,000점까지의 신용점수를 산정해 제공하고 있다(금융위원회, 2020). 개인신용평가에 인공지능이 도입되면서 평가 소요 시간을 단축하고 인건비 및 사용자의 수수료 부담을 낮출 수 있게 되었으며, 기존 신용평가모형에 비해 예측력과 변별력이 높아졌다(Yusof Ishak Institute, 2021; World Bank Group, 2019).

인공지능이 도입된 개인신용평가의 가장 큰 특성은 비금융정보(non-financial)의 활용이다(최성민, 2020). 비금융정보란 통신 및 전기요금 납부 이력, 온라인 쇼핑 적립 포인트, 소셜미디어(SNS) 사용 내용 등 신용도와 연관성이 낮은 온라인 및 오프라인의 모든 데이터를 포함하며, 비전통적(non-traditional) 또는 대안(alternative) 정

〈표 3〉 국내 개인신용평가모형에 활용되는 비금융정보

| 구분 | 회사 | 비금융정보 |
|---|---|---|
| 카드사 | 신한카드 | 휴대전화 단말기 가격, 해외 로밍 횟수, 할인 쿠폰 조회 수 |
| 핀테크 | 8 퍼센트 | 계약 단계별 사용자 체류 시간 및 클릭의 정확도, 휴대전화 주 사용 시간대 |
| | 크레파스 | SNS 의 소통 주기, 휴대전화 배터리 충전 주기, 문자메시지나 통화의 응답 반응 속도 |
| 인터넷 은행 | 카카오뱅크 | 카카오 T 택시 탑승 이력, 카카오 선물하기 사용 이력, 통신비 정상 납부 개월 수 |
| | 토스뱅크 | 월세 연체 여부, 적금 가입 여부, 고객 신용카드 사용 내용, 아르바이트 이력 |

보라고 부르기도 한다(ICCR, 2018; 권영준 등, 2011).

비금융정보를 반영한 개인신용평가모형에 대한 기초 연구를 진행한 PERC (2006)은 개인신용평가모형에 전기·가스 정보와 통신정보를 반영했을 때 개인신용평가모형의 예측률이 크게 향상했으며, 신용대출 연체율이 감소함을 밝혔다. 이와 비슷하게 국내에서도 전기요금 정보를 반영한 개인신용평가모형의 예측률이 향상하고 신용대출 관련 불량 고객 수와 연체율을 감소시킨다는 연구가 발표된 바 있다(권영준 등, 2011).

이런 현상은 특히 사회초년생, 가정주부, 학생 등 금융 거래 내용이 부족한 '금융이력부족자(Thin Filer)'에게 더 효과적으로 나타났다(PERC, 2009). 이처럼 그간 평가 요소로 활용되지 못했던 비금융정보를 활용할 수 있게 됨에 따라 거래정보(판매량, 평균 판매 금액), 평판 관련 정보(처리 시간, 후기 및 불만 사항) 등 기업이 보유한 데이터를 반영해 자체 개인신용평가모형을 개발하는 기업이 늘고 있으며(BIS, 2019), 예시는 <표 3>과 같다. 해외의 대표 사례로는 SNS 계정 수, 계정의 사용 기간 등 소셜 네트워크상의 평판 정보를 기반으로 소액 대출을 제공하는 미국 회사 렌도(Lenddo)가 있다. 현재 필리핀과 콜롬비아에서 금융 소비자 1인당 400~800달러 상당의 대출이 이뤄지고 있으며 대출 상환율이 95%에 달한다(전종헌, 2020). 최근 국내에서도 비금융정보를 주로 활용하는 '대안신용평가'가 非신용평가로 떠오르면서 비금융정보를 반영하는 개인신용평가모형이 발표되고 있다(안소영, 2021).

이러한 개인신용평가모형은 EU가 분류한 AI 시스템 위험도의 네 단계 중 고위험(high risk)으로 구분된다(Cornacchia et al., 2021). 이는 개인신용평가가 소비자의 전반적인 금융 생활 영위에 미치는 직접적인 영향력과 신용평가 서비스 과정에 내재한 위험성 때문이다(European Commission, 2019). 사람이 직접 평가했던 전통적 개인신용평가 결과를 개인신용평가모형의 학습데이터로 활용할 경우, 특정 평가자의 의사결정을 학습해 모형의 편향성을 유발할 수 있다는 위험성이 있다. 또한, 알고리즘을 학습시키고 검증하는 단계에서는 인공지능의 블랙박스 특성 때문에 개별 변수의 영향력을 확인하기 어렵다는 한계가 있다(Bruckner, 2018). 마지막으로 개인신용평가모형을 배포 및 운영하는 단계에서는 비금융정보의 활용으로 인한 개인정보 유출이나 프라이버시 침해의 위험성이 있고 특정 사용자 그룹에 대한 직·간접적인 차별로 이어질 수 있다(안소영, 2021). 따라서 AI 기반 개인신용평가의

공정성을 개선하기 위해서는 서비스 전 과정(기획부터 설계, 운영 및 모니터링)에 걸쳐 차별이나 편향 같은 비합리적인 요소가 반영되지 않도록 면밀한 검토가 필요하다.

### 3.2 공정한 AI 기반 개인신용평가를 위한 체크리스트 제안

공정성을 저해시키는 요인은 크게 편향(Bias)과 차별(Discrimination)로 구분된다(Mehrabi et al., 2021). 차별은 의사결정을 내리는 사람이 속해있는 사회 또는 환경에 의해 무의식적으로 학습되는 편견이나 심리적 고정관념으로부터 발생한다(황용석 등, 2021). 이에 비해 편향은 데이터 수집이나 데이터 샘플링 등 기술적 관점에서 모델을 활용하는 과정 중에 발생한다(Kozodoi et al., 2022). 본 연구는 연구 규모 및 범위를 고려하여 사회심리적인 측면에서 더 광범위한 연구가 필요한 차별요인을 제외하고 편향요인을 집중적으로 분석하였다.

편향요인은 Suresh and Guttag (2021)의 머신러닝 라이프사이클에 따른 편향과 Mehrabi et al. (2021)이 제안한 데이터-알고리즘-사용자 상호작용 루프 모형에 따른 편향을 기반으로 데이터, 알고리즘, 사용자 영역에 대한 9가지 편향을 <표 4>와 같이 재구성하였다. 9가지 편향은 순위 편향(Ranking bias), 표현 편향(Presentation bias) 등 개인신용평가와 연관성이 낮은 편향을 제외하고 AI 기반 개인신용평가의 특성을 반영하는 편향을 중심으로 선정하였다.

본 연구에서는 <표 4>을 개념적 프레임워크로 하여 개인신용평가와 관련된 연구논문, 관련 자

〈표 4〉 인공지능 기반 서비스에서 발생할 수 있는 데이터, 알고리즘, 사용자 영역의 편향

| 영역 | 편향 | 설명 | |
|---|---|---|---|
| Data | Representation Bias (대표 편향) | 설명 | 모집단에서 추출된 표본 집단이 모집단의 특성을 대표하지 못해 발생하는 편향이다. |
| | | 예시 | 스마트폰을 통해 수집된 데이터셋은 저소득 및 고령자 인구 통계를 제대로 포함하지 못할 가능성이 큼 |
| | Sampling Bias (표본 편향) | 설명 | 작위적으로 선정된 표본이 모집단을 대표하지 못할 때 발생하는 편향으로 결과의 일반화를 발생시킬 수 있다. |
| | | 예시 | 아마존의 얼굴 인식 시스템의 학습 데이터셋에 백인의 비율이 높아 유색인종에 대한 성능이 떨어져 논란이 됨 |
| | Measurement Bias (측정 편향) | 설명 | 모델에서 학습시키고자 하는 특성을 측정할 수 있는 요소로 정의하고 선정하는 과정에서 발생하는 편향이다. |
| | | 예시 | 과거 질병 이력, 처방 약물 등의 데이터를 통해 고위험군의 환자를 분류하고 예상 의료비용을 예측할 때, 비용이 높게 측정되는 환자를 고위험 환자로 분류하는 오류가 발생할 수 있음 |
| Algorithm | Aggregation Bias (집계 편향) | 설명 | 개인 또는 집단의 특성이 충분히 고려되지 않은 채 학습되어 다수 집단에 대해서만 정상적으로 수행될 때 발생하는 편향이다. |
| | | 예시 | 라틴 아메리카계 사람들은 백인들보다 당뇨병과 관련된 합병증의 비율이 더 높아서 당뇨병을 진단하는 AI 시스템을 구축하는 경우 데이터에 특정 민족성이 반영되도록 하거나 모델의 활용 대상을 구분하여 구축해야 함 |

| 영역 | 편향 | 설명 | |
|---|---|---|---|
| Algorithm | Learning Bias (학습 편향) | 설명 | 모델의 학습 과정에서 테스트 데이터셋에 따라 모델의 성능 격차가 발생하는 편향으로 모델의 목적에 따라 절충 요소가 발생한다. |
| | | 예시 | Differentially private training 방식이 학습데이터의 개인 정보 보호를 강화하지만, 모델의 성능 저하를 초래한다는 연구 발표(Bagdasaryan et al, 2019) |
| | Evaluation Bias (평가 편향) | 설명 | 모델을 평가하는 과정에서 발생하는 편향으로, 사용된 테스트 데이터셋이 모집단의 특성을 충분히 반영하지 못하여 발생한다. |
| | | 예시 | 대표적인 얼굴 분석 벤치마킹 데이터셋(Adience 및 IJB-A 등)이 밝은 피부의 모집단 위주로 구성되어 유색인종에 대한 성능이 현저히 떨어져 문제 제기(Buolamwini & Gebru, 2018) |
| User | Deployment Bias (배포 편향) | 설명 | 배포된 모델이 해결하고자 하는 문제와 실제 사용되는 방식과 다를 때 발생한다. 최종 사용자가 의도한 대로 모델을 사용하지 않는다면 모델이 잘 작동하지 않거나 의도되지 않은 오류를 발생시킬 수 있다. |
| | | 예시 | 범죄자의 재범률을 예측하는 시스템의 결과는 참고 자료로써만 활용되어야 하며, 판사가 적절한 처벌을 결정할 때 직접적으로 영향을 끼치는 자료로 활용해서는 안 됨 |
| | Historical Bias (역사 편향) | 설명 | 표본 데이터의 무결성과 관계없이 이미 과거부터 사회에 존재하고 있는 편향으로 사회적 고정관념 등으로 인해 특정 그룹에 불리하게 적용될 때 주로 발생한다. |
| | | 예시 | 구글의 뉴스 기사를 통해 학습된 모델이 성별에 대한 고정관념을 내포하고 이를 영구화 한다는 문제 제기(Bolukbasi et al., 2016) |
| | Social Bias (사회적 편향) | 설명 | 모델이 도출한 결과에 대한 타인의 반응 및 행동이 사용자의 판단에 영향을 끼치면서 발생하는 편향이다. |
| | | 예시 | 작품에 대한 평가를 진행할 때, 주변 사람들의 평가에 영향을 받아 더 낮거나 더 높게 평가하는 상황(Baeza-Yates, 2018) |

료 등의 내용을 분석하고 AI 기반 개인신용평가의 공정성 개선을 위한 데이터, 알고리즘, 사용자 영역에 대한 체크리스트를 구성하였다. 각 점검 항목은 AI 기반 개인신용평가의 공정성과 관련한 자료를 그룹화하고 영역별 편향과 매칭해 항목화하였다. 최종적으로 제안한 체크리스트는 3개의 영역과 12개의 점검 항목 및 항목별 세부 권고안으로 구성된다.

<표 5>는 데이터 영역에 대한 체크리스트로 알고리즘의 품질에 직접적인 영향을 끼치는 데이터를 설계·수집·가공하는 과정에서 발생할 수 있는 편향에 대한 점검 항목으로 구성하였다. 이

〈표 5〉 AI 기반 개인신용평가의 공정성 개선 체크리스트 – 데이터 영역

| 영역 | 점검 항목 & 적용 편향 | | | 항목별 세부 권고안 | 참고문헌 |
|---|---|---|---|---|---|
| 데이터 | 1-1 | 평가 항목의 적합성 검증 | Measurement Bias | • 서비스의 궁극적 목적 및 타겟을 고려해 신용평가의 평가 요소를 선정한다.<br>• 추상적인 개념의 평가 요소는 표본 집단에서 측정할 수 있는 요소로 정의해야 한다.<br>• 선정한 평가 요소와 신용도의 상관관계를 명확히 이해하고 평가 요소 간의 우선순위를 도출해야 한다. | (Suresh & Guttag, 2021) |

| 영역 | 점검 항목 & 적용 편향 | | | 항목별 세부 권고안 | 참고문헌 |
|---|---|---|---|---|---|
| 데이터 | 1-2 | 비금융정보 활용의 타당성 검증 | Measurement Bias | • CB 사가 제공하는 신용점수 외에 비금융정보를 활용할 경우, 신용도와의 상관관계를 확인해야 한다.<br>• 수집하고자 하는 비금융정보가 사용자의 프라이버시를 침해하지 않도록 특히 유의해야 하며, 데이터의 조합이 개인 또는 그룹을 특정할 수 있는 정보로 작용하지 않도록 모니터링이 필요하다.<br>• 수집하고자 하는 비금융정보가 어느 표본 집단에서든 동일한 조건으로 측정 가능한지 확인해야 한다. | (European Commission, 2018), (World Bank, 2018), (World Bank and CGAP, 2018) |
| | 1-3 | 기 구축된 학습 데이터셋에 내재한 편향성 확인 | Measurement Bias | • 사람의 의사결정에 기반한 전통적 신용평가 결과를 인공지능 모델에 학습시킬 경우, 기존 데이터에 내재한 편향성을 확인해야 한다.<br>• 학습 데이터셋의 최적화를 위해 새로 수집이 필요한 데이터와 배제할 데이터를 구분해야 한다. | (World Bank Group, 2019) |
| | 1-4 | 표본 집단 설계의 타당성 검증 | Representation Bias & Sampling Bias | • 서비스 타겟의 특성을 고려해 표본 집단을 정의하고 균형 있는 데이터 수집을 위한 표본 집단의 비율을 설정해야 한다.<br>• 타겟 및 표본 집단에서 소외되는 집단이 없는지 확인해야 한다.<br>• 학습 데이터셋과 테스트 데이터셋은 위의 과정을 고려해 동일한 방식으로 구축되어야 한다. | (Mehrabi et al., 2021), (Petrasic et al., 2017) |

〈표 6〉 AI 기반 개인신용평가의 공정성 개선 체크리스트 – 알고리즘 영역

| 영역 | 점검 항목 & 적용 편향 | | | 항목별 세부 권고안 | 참고문헌 |
|---|---|---|---|---|---|
| 알고리즘 | 2-1 | 목적에 맞는 성능 평가지표의 최적화 | Evaluation Bias | • 알고리즘 최적화를 위해 다방면을 고려한 성능 평가지표(Evaluating Metric) 구축이 필요하다. 예를 들어, 신용점수를 활용한 대출 심사 서비스에서 관점(사용자: 실제 받을 수 있는 조건이지만 거절당하는 상황 우려, 금융회사: 대출을 갚지 못하는 사람에게 대출을 승인하는 상황 우려)에 따라 서비스를 제공하는 방식에 차이가 있을 수 있다.<br>• 충분히 다양한 테스트셋을 활용해 모델을 검증하고, 신용평가 결과를 설명할 수 있는 사례를 확보해야 한다. | (FSB, 2017), (Petrasic et al., 2017) |
| | 2-2 | 외부 알고리즘 적용의 적합성 확인 | Aggregation Bias & Evaluation Bias | • 외부 알고리즘을 도입하려고 할 경우, 해당 신용평가모형 구축 목적에 부합하는지 충분한 고려가 필요하다. 특히 해외 알고리즘의 적용은 문화적 및 사회적 차이로 인하여 예상하지 못한 결과를 유발할 수 있음을 유의해야 한다.<br>• 도입하고자 하는 알고리즘에 내포된 편향은 다양한 테스트셋을 통해 적합성 및 공정성을 검증해야 한다. | (엄하늘 등, 2020), (Ntoutsi et al., 2020), (World Bank Group, 2019) |
| | 2-3 | 데이터에 대한 모델 성능 동일성 검증 | Learning Bias | • 특정 데이터에만 과적합 되어 데이터셋 별로 평가 결과가 상이하게 나타나는 경우가 없도록 유의해야 한다.<br>• 모델의 학습 방향이 특정 데이터를 해석하기에 용이한 형태로 선정되지 않도록 점검해야 한다.<br>• 특성이 다른 데이터로 모델을 평가할 시, 데이터별 참작이 가능한 성능 범위를 정하여 점검해야 한다. | (Petrasic et al., 2017), (World Bank Group, 2019) |
| | 2-4 | 지속적인 모니터링을 통한 신규 유입 데이터 편향성 확인 | Aggregation Bias | • 신용평가모형이 배포된 후, 축적되는 사용자 데이터를 학습 데이터셋으로 활용하기 위해서는 체계적인 설계가 필요하다.<br>• 수집된 신용평가 결과를 실시간으로 학습시킬 경우, 특정 집단의 데이터 비중이 커져 편향을 발생시키지 않도록 유의해야 한다. | (Mehrabi et al., 2021 |

〈표 7〉 AI 기반 개인신용평가의 공정성 개선 체크리스트 – 사용자 영역

| 영역 | 점검 항목 & 적용 편향 | | | 항목별 세부 권고안 | 참고문헌 |
|---|---|---|---|---|---|
| 사용자 | 3-1 | 신용평가모형의 영향력 및 사용자 반응 모니터링 | Social Bias & Deployment Bias | • 신용평가 결과가 사용자에게 공정한 수준으로 적용되는지에 대한 검토가 필요하다.<br>• 신용평가 결과로 발생할 수 있는 오해 및 편향을 사전에 방지하기 위해 심사 결과가 사용자에게 미치는 영향을 확인하고, 사용자가 특정 평가 요소에 대해 얼마나 민감하게 반응하는지 등의 사용자 반응에 대한 검토가 필요하다. | (Suresh & Guttag, 2019), (World Bank Group, 2019) |
| | 3-2 | 결과에 대한 해석 오류 및 편향성 확인 | Social Bias & Deployment Bias | • 사용자가 신용평가모형이 도출한 결과를 잘못된 기준으로 인식하지 않도록 확인해야 한다.<br>• 결과에 대한 잘못된 인식으로 사용자가 의도적으로 평가 항목(특히 비금융정보)의 조건에 맞추는 인위적인 행동을 유발하지 않도록 유의해야 한다.<br>• 사용자가 요청할 경우, 신용평가 결과의 공정성에 대한 설명을 제공할 수 있어야 한다. | (König-Kersting et al,. 2021) |
| | 3-3 | 신용평가모형의 오용으로 인한 편향성 확인 | Deployment Bias | • 신용평가 결과는 참고용 자료로 추가적인 의사결정 과정이 필요하며, 최종 판단의 도구로 활용되는 상황을 유의해야 한다.<br>• 사용자에게 신용평가 결과에 대한 활용 가이드를 제공하여 사용자가 정확한 신용평가의 목적을 이해하도록 도와야 한다. | (World Bank Group, 2019) |
| | 3-4 | 사회적 요인의 변화에 따른 지속적 유지보수 | Social Bias & Historical Bias | • 신용평가모형의 개발 중 예상하지 못해 발생한 변수에 대한 지속적인 모니터링이 필요하다.<br>• 문화적 및 환경적 변화에 따라 추가로 확장되는 데이터에 대한 유지보수가 진행되어야 한다. | (Ntoutsi et al., 2020), (Suresh & Guttag, 2019) |

영역에서는 데이터 설계자가 평가 항목 선정에 대한 타당성을 명확하게 정의하고, 데이터를 설계 및 수집하는 과정 중에 소외되는 집단이 없도록 면밀하게 점검해야 한다.

<표 6>은 알고리즘 영역에 대한 체크리스트로 개인신용평가모형을 구축·학습·검증하는 과정에서 발생할 수 있는 편향을 중심으로 구성하였다. 이 영역에서 모델 개발자는 '정확도'와 '공정성' 같은 상충관계를 신중하게 고려하고 어떤 데이터로 검증하더라도 모든 표본 집단이 수용할 수 있는 결과를 도출할 수 있도록 평가지표를 최적화해야 한다.

<표 7>은 사용자 영역에 대한 체크리스트로 개인신용평가모형을 배포 및 운영하는 단계에서 발생할 수 있는 편향을 중심으로 구성하였다. 이 영역에서 서비스 운영자는 신용평가 결과가 사용자와 데이터에 미치는 영향력을 관찰하여 발생할 수 있는 편향과 오류를 방지하고자 하는 노력이 필요하다.

# 4. 제안한 체크리스트에 대한 AHP 분석

## 4.1 분석 모델 설계

본 연구는 AI 기반 개인신용평가의 공정성에 대한 여러 관점의 의견을 합리적으로 포용하기 위해 금융 분야 종사자, 인공지능 분야 종사자, 일반 사용자로 분류한 이해관계자를 대상으로

〈표 8〉 설문 응답자 특성

| 특성 | 구분 | 전체 | 이해관계자별 | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | 금융 분야 종사자 | | 인공지능 분야 종사자 | | 일반 사용자 | |
| 성별 | 남자<br>여자 | 14 명<br>13 명 | 7 명<br>1 명 | | 5 명<br>4 명 | | 2 명<br>8 명 | |
| 연령 | 20 대<br>30 대<br>40 대 이상 | 9 명<br>14 명<br>4 명 | 1 명<br>3 명<br>4 명 | | 3 명<br>6 명<br>0 명 | | 5 명<br>5 명<br>0 명 | |
| 종사 분야 / 직업 | | | IT / 전산<br>소비자 및 데이터<br>핀테크 서비스<br>창구 및 출납 | 2 명<br>2 명<br>3 명<br>1 명 | 모델 개발 및 연구<br>데이터 수집 | 8 명<br>1 명 | 대학원생<br>회사원<br>전문직<br>무직 | 6 명<br>2 명<br>1 명<br>1 명 |
| 종사 경력 | 1~5 년<br>6~10 년<br>10 년 이상 | 11 명<br>3 명<br>3 명 | 3 명<br>2 명<br>3 명 | | 8 명<br>1 명<br>0 명 | | - | |

AHP 분석을 진행하였다. AHP는 목표 달성을 위해 다수의 기준 간 상대적 중요도와 우선순위를 산정하는 기법으로 의사결정 과정에서 정량적 요소와 정성적인 요소를 동시에 객관화하고 체계화하는 방안으로 활용된다(Saaty, 1994). 특히 이해관계자를 금융서비스의 특성을 잘 이해하는 금융 분야 종사자, AI 모델 구축과 관련해 전문적 배경지식이 있는 인공지능 분야 종사자, AI 기반 개인신용평가 서비스의 잠재적 고객인 일반사용자로 구분하고 인식 차이를 확인함을 통해 사용자 친화적인 AI 기반 개인신용평가 설계에 도움을 줄 것으로 판단하였다. 설문은 체크리스트에서 세부 권고안을 제외한 3개 영역과 12개 점검 항목을 클라우드 사회과학 연구 자동화(Social Science Research Automation, SSRA)가 제공하는 9점 척도 쌍대비교(pairwise comparison) 방식으로 구성하였다. 설문은 총 27명(금융 분야 종사자 8명, 인공지능 분야 종사자 9명, 일반 사용자 10명)이 참여하였고 응답자 특성은 <표 8>과 같다. 이창효 (2000)에 따르면 실무지식이나 전문적 경험이 있을 때 AHP의 표본 크기를 10~15명 내외로 하는 것으로 알려져 있다. 본 연구에서는 설문 참여자가 현업 종사자이면서 관련 서비스에 대한 이해(금융 분야 종사자-AI 시스템, 인공지능 분야 종사자-AI 기반 개인신용평가)가 있어야 한다는 제약으로 설문 참여자 확보에 어려움이 있어 전문가 그룹인 금융 분야 종사자와 인공지능 분야 종사자의 표본 크기가 권고안보다 다소 작았지만, AHP 결과의 신뢰성을 보장할 수 있는 수준의 일관성 비율은 확보하였다.

### 4.2 결과

AHP 결과의 신뢰성 확보를 위해서는 일관성 검증이 필수적이다(Saaty, 1994). 이를 위해 일관성 지수(Consistency Index, CI)를 임의 지수(Random Index, RI)로 나눠 계산한 일관성 비율(Consistency Ratio, CR)을 도출하고 일관성 비율이 0.1 이하일 때 설문 응답을 신뢰할 수 있다고 판단한다(Saaty, 1990). 본 연구의 전체 CR은 0.00157이며, 각 영역의 일관성 비율은 데이터(0.00290), 알고리즘(0.00178), 사용자(0.00210)로

〈표 9〉 영역별 상대적 중요도 비교

| 영역 | 전체 | 금융 분야 종사자 | 인공지능 분야 종사자 | 일반 사용자 |
|---|---|---|---|---|
| 데이터 | 0.466 | 0.406 | 0.586 | 0.418 |
| 알고리즘 | 0.317 | 0.375 | 0.232 | 0.343 |
| 사용자 | 0.217 | 0.220 | 0.182 | 0.239 |

〈표 10〉 통합 중요도를 적용한 이해관계자별 점검 항목 중요도 및 우선순위

| 영역 | 중요도 (A) | 점검 항목 | 전체 | | | 금융 분야 종사자 | | 인공지능 분야 종사자 | | 일반 사용자 | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | 영역별 중요도(B) | 통합 중요도 (A*B) | 통합 우선순위 | 중요도 | 우선순위 | 중요도 | 우선순위 | 중요도 | 우선순위 |
| 데이터 | 0.466 | 1-1 | 0.294 | 0.137 | 1 | 0.165 | 1 | 0.127 | 3 | 0.118 | 1 |
| | | 1-2 | 0.197 | 0.092 | 5 | 0.079 | 6 | 0.111 | 4 | 0.082 | 8 |
| | | 1-3 | 0.258 | 0.120 | 2 | 0.066 | 8 | 0.211 | 1 | 0.108 | 3 |
| | | 1-4 | 0.250 | 0.117 | 3 | 0.096 | 4 | 0.137 | 2 | 0.109 | 2 |
| 알고리즘 | 0.317 | 2-1 | 0.271 | 0.086 | 7 | 0.130 | 2 | 0.053 | 9 | 0.083 | 7 |
| | | 2-2 | 0.145 | 0.046 | 10 | 0.048 | 10 | 0.025 | 12 | 0.067 | 10 |
| | | 2-3 | 0.299 | 0.095 | 4 | 0.114 | 3 | 0.060 | 7 | 0.107 | 4 |
| | | 2-4 | 0.285 | 0.090 | 6 | 0.083 | 5 | 0.094 | 5 | 0.086 | 6 |
| 사용자 | 0.217 | 3-1 | 0.155 | 0.034 | 12 | 0.044 | 11 | 0.027 | 11 | 0.031 | 12 |
| | | 3-2 | 0.347 | 0.075 | 8 | 0.076 | 7 | 0.058 | 8 | 0.088 | 5 |
| | | 3-3 | 0.300 | 0.065 | 9 | 0.058 | 9 | 0.061 | 6 | 0.072 | 9 |
| | | 3-4 | 0.198 | 0.043 | 11 | 0.042 | 12 | 0.036 | 10 | 0.049 | 11 |

설문 결과에 대한 신뢰성을 확보하였다.

각 이해관계자 그룹의 영역별 중요도는 <표 9>와 같다. 전체 그룹의 우선순위는 '데이터(0.466)', '알고리즘(0.317)', '사용자(0.217)' 순으로 나타났다. 이해관계자별 중요도는 전체 그룹의 결과와 비슷하지만, 인공지능 분야 종사자의 데이터 영역 중요도(0.586)가 다른 영역에 비해 높게 나타났다. 점검 항목의 통합 중요도는 각 영역의 중요도와 영역별 점검 항목의 중요도를 통해 <표 10>과 같이 산출하였으며, 통합 중요도를 적용한 점검 항목의 우선순위는 <그림 1>과 같다. 이해관계자별로 중요도 차이가 적은 항목은 2-4, 3-4, 3-3 순으로 나타났으며 대부분 사용자 영역에 속한다는 공통점이 있다. 한편 이해관계자별로 뚜렷한 차이를 보인 점검 항목은 1-3, 2-1, 2-3으로 특히 금융 분야 및 인공지능 분야 종사자의 차이가 두드러지게 나타났다.


〈그림 1〉 영역별 점검 항목의 우선순위를 반영한 계층도

### 4.3 논의

본 연구에서는 현업에서 활용 중인 AI 기반 개인신용평가 서비스와 관련해 연구 결과가 가지는 실용적인 시사점을 논의하고자 하였다. 다양한 관점을 수용하고 상대적 중요도를 유의미하게 활용하기 위해 이해관계자 그룹별로 비슷하거나 차이를 보이는 점검 항목을 통합 중요도의 평균값인 0.083을 기준으로 세 가지 유형으로 분류해 분석을 진행하였다.

◆ 제1유형 : 모든 그룹에서 0.083 이상의 중요도를 받은 점검 항목
◆ 제2유형 : 한 그룹을 제외하고 0.083 이상의 중요도를 받은 점검 항목
◆ 제3유형 : 모든 그룹에서 0.083 미만의 중요도를 받은 점검 항목

<표 11>의 점검 항목인 1-1과 1-4는 모든 이해관계자가 가장 큰 관심을 나타낸 부분이다. 두 항목의 공통점은 데이터를 설계하는 단계에서 고려되어야 하는 요소이면서 신용평가 결과에 가장 직접적인 영향을 미치는 기준으로 작용한다는 특징이 있다. 따라서 개인신용평가모형을 위한 데이터를 설계하는 과정에서 중요하게 다뤄지고 있는지에 대한 점검이 필요하다.

〈표 11〉 제1유형 점검 항목의 상대적 중요도

| 영역 | 점검 항목 | 전체 | 금융 분야 종사자 | 인공지능 분야 종사자 | 일반 사용자 |
|---|---|---|---|---|---|
| 데이터 | 1-1 | 0.137 | 0.165 | 0.127 | 0.118 |
| | 1-2 | 0.092 | 0.079 | 0.111 | 0.082 |
| | 1-4 | 0.117 | 0.096 | 0.137 | 0.109 |
| 알고리즘 | 2-4 | 0.090 | 0.083 | 0.094 | 0.086 |

<표 11>의 점검 항목 1-2에 대한 금융 분야 종사자와 일반 사용자의 상대적 중요도가 기준점(0.083)에서 떨어지지만, 그 차이가 미세하고 현재 국내에서 비금융정보를 활용하는 대안신용평

가가 떠오르고 있는 만큼 모든 이해관계자가 관심을 가지고 주목한 부분으로 해석할 수 있다. 비금융정보는 대출 신청 시 클릭의 정확도, 월세 연체 여부 등 신용도와 연관성이 적은 항목들이 대부분이기 때문에 이에 대한 타당성 검증이 중요하게 고려되어야 할 필요가 있다.

<표 11>의 점검 항목 2-4는 모든 이해관계자가 주목하였으며 신규 유입 데이터의 모니터링을 고려한 점검 항목이다. 이는 모든 이해관계자가 추가되는 데이터를 계속 학습해가는 머신러닝의 특성을 이해하고 있다는 것을 설명하고, 개인신용평가모형의 배포 이후에도 지속적인 모니터링이 중요하게 고려가 필요할 것으로 해석할 수 있다. 따라서 개인신용평가모형이 모든 사용자에게 공정한 결과를 제공하기 위해 서비스의 모든 과정에서 데이터의 편향성을 세심하게 살피는 지속적인 모니터링이 필요하다.

〈표 12〉 제2유형 점검 항목의 상대적 중요도

| 영역 | 점검 항목 | 전체 | 금융 분야 종사자 | 인공지능 분야종사자 | 일반 사용자 |
|---|---|---|---|---|---|
| 데이터 | 1-3 | 0.120 | 0.066 | 0.211 | 0.108 |
| 알고리즘 | 2-1 | 0.086 | 0.130 | 0.053 | 0.083 |
| | 2-3 | 0.095 | 0.114 | 0.060 | 0.107 |

<표 12>의 점검 항목 1-3에 대해 인공지능 분야 종사자는 높은 관심을 보였지만 금융 분야 종사자는 크게 주목하지 않았다. 이는 전통적 신용평가방식에 대한 금융 분야 종사자의 높은 신뢰가 반영된 것으로 해석할 수 있으며, 개인신용평가모형에 대한 신뢰도가 전통적 방식과 차이가 있을 수 있음을 내포한다. 인공지능 도입에 따른 신용평가 방식 또는 개인신용평가모형의 종류에 따른 사용자 신뢰도 변화를 확인하는 연구가 필요하다.

<표 12>의 점검 항목인 2-1과 2-3은 인공지능 분야 종사자를 제외한 이해관계자가 유의미한 관심을 보였는데 특히 금융 분야 종사자가 높이 평가했다. 이러한 현상은 각 이해관계자의 업무적 특성에 기인하여 발생한 현상으로 해석할 수 있으며, 추후 연구에서 각 이해관계자의 사용 경험 분석을 통한 검증이 필요하다.

〈표 13〉 제3유형 점검 항목의 상대적 중요도

| 영역 | 점검 항목 | 전체 | 금융 분야 종사자 | 인공지능 분야종사자 | 일반 사용자 |
|---|---|---|---|---|---|
| 알고리즘 | 2-2 | 0.046 | 0.048 | 0.025 | 0.067 |
| 사용자 | 3-1 | 0.034 | 0.044 | 0.027 | 0.031 |
| | 3-2 | 0.075 | 0.076 | 0.058 | 0.088 |
| | 3-3 | 0.065 | 0.058 | 0.061 | 0.072 |
| | 3-4 | 0.043 | 0.042 | 0.036 | 0.049 |

<표 13>의 점검 항목인 3-1과 3-4는 서비스가 배포된 이후에 확인할 수 있다는 공통점이 있다. 배포 이후의 점검 항목에 대한 관심도가 낮은 것은 역으로 제1유형에 나타난 것과 같이 모든 이해관계자가 배포 이후의 과정보다 모델 설계의 초기 과정을 더욱 중요하게 고려하고 있음을 한 번 더 확인시켜준다.

흥미롭게도 <표 13>의 점검 항목 3-2에 대해 일반 사용자만 높은 관심을 보였는데 이 패턴이 점검 항목 3-3에서도 비슷하게 나타났다. 이는 한도 결정, 대출 승인 등의 의사결정 시 신용점수가 중요하게 작용한다는 배경하에 AI 시스템으로 인한 불이익을 받지 않을까 하는 사용자의 불안 심리가 작용한 것으로 해석할 수도 있으며,

이 부분에 관해서는 향후 연구를 통한 추가 검증이 필요하다.

## 5. 결론

본 연구에서는 공정한 개인신용평가모형의 구축과 관리 과정을 점검하기 위해 활용할 수 있는 체크리스트를 제안하였다. 체크리스트는 문헌연구를 바탕으로 데이터, 알고리즘, 사용자의 3개 영역으로 구분하여 12가지 점검 항목과 항목별 세부 권고안으로 구성하였다. 체크리스트의 점검 항목별 상대적 중요도와 관련 이해관계자들의 인식 차이를 확인하기 위해 AHP 분석을 진행하였으며, 분석 결과를 기반으로 실제 신용평가 서비스의 전 과정을 고려하여 시사점을 도출하였다.

본 연구의 의의는 다음과 같다. 첫째, 체크리스트를 AI 기반 개인신용평가 서비스의 기획, 설계, 운영, 모니터링까지의 전 과정을 고려해 체계적으로 제안하였다. 둘째, AHP 분석은 이해관계자를 세분화해 진행함을 통해 AI 기반 개인신용평가에 대한 인식 차이를 확인하고 다양한 관점을 수렴해 사용자 친화적인 서비스 구축에 기여하고자 하였다. 셋째, 현업에서 활용 중인 AI 기반 개인신용평가 서비스에 관련해 연구 결과가 가지는 실용적인 시사점을 구체적으로 논의하였다. 마지막으로 제안한 체크리스트가 신뢰할 수 있는 인공지능 기반 금융서비스 구축을 위해 모델을 개선하는 정량적 방식 이외에 정성적으로 참고할 수 있는 기초자료로 유용하게 활용될 수 있을 것으로 기대한다.

본 연구는 학술적, 실무적 의의와는 별개로 다음과 같은 한계점이 있다. 첫째, 연구 규모와 범위를 고려하여 AI 시스템의 공정성을 저해시키는 요인인 편향과 차별 중에서도 편향요인에 대한 분석이 집중적으로 이뤄졌다. 따라서 AI 시스템의 공정성에 대한 전체적인 측면을 고려하지 못했다는 한계가 있다. 둘째, 연구 대상이 국내의 AI 기반 개인신용평가이지만 실제 산업 환경을 고려한 조사 없이 문헌을 중심으로 참고하였다는 한계점이 있다.

향후 연구에서는 현장 조사를 통해 운용 중인 AI 기반 개인신용평가 서비스의 현황을 파악하고 국내 금융서비스의 특성을 고려하여 체크리스트의 효용성을 검증하고자 한다. 또한, 실제 서비스에 대한 소비자의 신뢰 및 인식을 확인하고 설명 가능한 인공지능 관점에서의 고객 경험 개선안을 연구하고자 한다.

## 참고문헌(References)

**[국내 문헌]**

과학기술정보통신부. (2021). 사람이 중심이 되는 인공지능을 위한 신뢰할 수 있는 인공지능 실현 전략[안]. Retrieved 2022년 3월 12일, from https://www.korea.kr/common/download.do?fileId=195009613&tblKey=GMN

권영준, 남재현, 조민정. (2011). 개인신용평가에서의 비금융정보의 경제적 효과. *한국경제연구*, *29*(2), 81-107.

금융위원회. (2020년). '21.1.1일부터는 신용점수로 자신의 신용을 확인하세요. Retrieved 2022년 3월 12일, from http://www.fsc.go.kr:8300/v/p42S1u6Twh2

금융위원회. (2021). 금융분야 AI 가이드라인 및 주요 검토 필요사항. Retrieved 2022년 3월 5일, from http://www.fsc.go.kr:8300/v/pq8TQ

UQFZSY

금융위원회. (2021). 코로나 이후 시대의 디지털 대전환을 선도하기 위해 금융분야 인공지능(AI)을 활성화하겠습니다. Retrieved 2022년 4월 6일 from http://www.fsc.go.kr:8300/v/pbVFpTRt0h5

김지웅, 허준, 김장일. (2013). 빅데이터의 금융기관 활용 사례. *The Magazine of the IEIE*, *40*(8), 49 - 54.

소순주, 안성진. (2021). 인공지능 윤리원칙 분류 모형 및 구성요소에 관한 연구. *컴퓨터교육학회 논문지*, *24*(6), 119-132.

안소영. (2021년, 9월 16일). 대안신용평가 시대 온다… 전통 금융사들 빅데이터 센터 세워 대응해야. ChosunBiz. https://biz.chosun.com/stock/finance/2021/09/16/7MJQZ34FLFD7RHSFY7D3TR342A/

양희태, 최병삼, 이제영, 장훈, 백서인, 김단비. (2018). 인공지능 기술 전망과 혁신정책 방향 - 국가 인공지능 R&D 정책 개선방안을 중심으로-. Retrieved 2022년 3월 2일, from https://www.nkis.re.kr:4445/researchReport_view.do?otpId=OTP_0000000000002198#none

엄하늘, 김재성, 최상옥. (2020). 머신러닝 기반 기업부도위험 예측모델 검증 및 정책적 제언: 스태킹 앙상블 모델을 통한 개선을 중심으로. *지능정보연구*, *26*(2), 105-129.

이정선, 서보밀, 권영옥. (2021). 인공지능이 의사결정에 미치는 영향에 관한 연구: 인간과 인공지능의 협업 및 의사결정자의 성격 특성을 중심으로. *지능정보연구*, *27*(3), 231-252.

이창효. (2000). *집단의사결정론*. 서울: 세종출판사.

장용석, 김형준, 문정욱, 문아람, 김정언, 이시직, 양기문, 황선영, 변순용, 선지원, 이청호, 김봉제. (2020). *윤리적 인공지능을 위한 국가 정책 수립*. *정책연구*, *2020*(7), 1-235.

전종헌. (2020년, 6월 4일). SNS활동 신용평가해 대출했더니 '상환율 95%'...비금융정보 주목. 매일경제. https://www.mk.co.kr/news/economy/view/2020/06/573304/

최성민. (2020). 개인 신용평가모형과 설명력 이슈. 2020 한국경영정보학회 추계학술대회, 한국과학기술회관, 서울.

황용석, 정재선, 황현정, 김형준. (2021). 알고리즘 추천 시스템의 공정성 확보를 위한 시론적 연구. *방송통신연구*, 169-206.

KDB 미래전략연구소. (2021). '금융분야 AI 가이드라인' 및 금융권의 대응. Retrieved 2022년 2월 26일, from https://eiec.kdi.re.kr/policy/domesticView.do?ac=0000159143

**[국외 문헌]**

Baeza-Yates, R. (2018). Bias on the web. *Communications of the ACM*, *61*(6), 54-61.

Bagdasaryan, E., Poursaeed, O., & Shmatikov, V. (2019). Differential privacy has disparate impact on model accuracy. 33rd Conference on Neural Information Processing System, Vancouver, Canada.

Bank for International Settlements(BIS) (2019). Big tech in finance: opportunities and risks. Retrieved December 26, 2021, from https://www.bis.org/publ/arpdf/ar2019e3.pdf

Bolukbasi, T., Chang, K. W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. 30th Conference on Neural Information Processing System, Barcelona, Spain.

Bruckner, M. A. (2018). Regulating fintech lending. Banking & Financial Services *Policy Report*, *37*(6).

Buolamwini, J., & Gebru, T. (2018, February). Gender shades: Intersectional accuracy disparities in commercial gender classification. ACM Conference on fairness, accountability and transparency, New York, USA.

Cornacchia, G., Narducci, F., & Ragone, A. (2021, September). A general model for fair and explainable recommendation in the loan domain. Joint Proceedings KaRS 2021 and ComplexRec 2021, Amsterdam, Netherlands.

European Commission. (2018). Communication Artificial Intelligence for Europe. Retrieved January 13, 2022, from https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52018DC0237&from=EN

European Commission. (2019). Ethics guidelines for trustworthy AI, Report. Retrieved January 13, 2022, from https://www.aepd.es/sites/default/files/2019-12/ai-ethics-guidelines.pdf

Financial Stability Board(FSB). (2017). Artificial Intelligence and Machine Learning in Financial Services: Market Developments and Financial Stability Implications. Retrieved March 23, 2022, from https://www.fsb.org/wp-content/uploads/P011117.pdf

Fuster, A., Goldsmith Pinkham, P., Ramadorai, T., & Walther, A. (2022). Predictably unequal? The effects of machine learning on credit markets. *The Journal of Finance, 77*(1), 5-47.

Hardt, M., Price, E., & Srebro, N. (2016, December). Equality of opportunity in supervised learning. 30th Conference on Neural Information Processing System, Vancouver, Canada.

International Committee on Credit Reporting (ICCR). (2018). Guidance Note: Use of Alternative Data to Enhance Credit Reporting to Enable Access to Digital Financial Services by Individuals and SMEs Operating in the Informal Economy. Retrieved January 24, 2022, from https://www.gpfi.org/sites/gpfi/files/documents/Use_of_Alternative_Data_to_Enhance_Credit_Reporting_to_Enable_Access_to_Digital_Financial_Services_ICCR.pdf

König-Kersting, C., Pollmann, M., Potters, J., & Trautmann, S. T. (2021). Good decision vs. good results: Outcome bias in the evaluation of financial agents. *Theory and Decision, 90*(1), 31-61.

Kozodoi, N., Jacob, J., & Lessmann, S. (2022). Fairness in credit scoring: Assessment, implementation and profit implications. *European Journal of Operational Research*, *297*(3), 1083-1094.

Lepri, B., Oliver, N., Letouzé, E., Pentland, A., & Vinck, P. (2018). Fair, transparent, and accountable algorithmic decision-making processes. *Philosophy & Technology*, *31*(4), 611-627.

Li, J., & Chignell, M. (2022). FMEA-AI: AI fairness impact assessment using failure mode and effects analysis. *AI and Ethics,* 1-14.

McCalman, L., Steinberg, D., Abuhamad, G., Brunet, M. E., Williamson, R. C., & Zemel, R. (2022). Assessing AI Fairness in Finance. *Computer, 55*(1), 94-97.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR), 54*(6), 1-35.

Microsoft. (2020). Fairlearn: A toolkit for assessing and improving fairness in AI. Retrieved April 2, 2022, from https://www.microsoft.com/en-us/research/uploads/prod/2020/05/Fairlearn_WhitePaper-2020-09-22.pdf

Monetary Authority of Singapore(MAS). (2020). FEAT Fairness Principles Assessment Case Studies. Retrieved April 25, 2022, from https://www.mas.gov.sg/-/media/MAS/News/Media-Releases/2021/Veritas-Document-2-FEAT-Fairness-Principles-Assessment-Case-Studies.pdf

Ntoutsi, E., Fafalios, P., Gadiraju, U., Iosifidis, V., Nejdl, W., Vidal, M. E., Ruggieri, S., Turini, F., Papadopoulous, S., Krasankis, E., Kompatsiaris, I., Kurlanda, K. K., Wagner, C., Karimi, F., Fernandez, M., Alani, H., Berendt, B., Kruegel, T., Heinze, C., Broelemann, K., Kasneci, G., Tiropanis, T., & Staab, S. (2020). Bias in data driven artificial intelligence systems－An introductory survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 10*(3).

Petrasic, K., Saul, B., Greig, J., & Bornfreund, M. (2017). Algorithms and bias: What lenders need to know. Retrieved April 25, 2022, from https://www.lexology.com/library/detail.aspx?g=c806d996-45c5-4c87-9d8a-a5cce3f8b5ff

Political & Economic Research Council(PERC). (2006). Give credit where credit is due: Increasing access to affordable mainstream credit using alternative data. Retrieved January 3, 2022, from https://www.brookings.edu/wp-content/uploads/2016/06/20061218_givecredit.pdf

Political & Economic Research Council(PERC). (2009). New to Credit from Alternative Data. Retrieved May 3, 2022, from https://www.perc.net/wp-content/uploads/2013/09/New_to_Credit_from_Alternative_Data_0.pdf

Saaty, T. L. (1990). How to make a decision: the analytic hierarchy process. *European journal of operational research, 48*(1), 9-26.

Saaty, T. L. (1994). Highlights and critical points in the theory and application of the analytic hierarchy process. *European journal of operational research, 74*(3), 426-447.

Suresh, H., & Guttag, J. (2021). A framework for understanding sources of harm throughout the machine learning life cycle. *In Equity and Access in Algorithms, Mechanisms, and Optimization,* 1-9.

Vigdor, N. (2019, November 10). Apple Card Investigated After Gender Discrimination Complaints. The New York Times. https://www.nytimes.com/2019/11/10/business/Apple-credit-card-investigation.html

World Bank Group (2018). Financial Consumer Protection and New Forms of Data Processing Beyond Credit Reporting. Retrieved April 23, 2022, from https://openknowledge.worldbank.org/bitstream/handle/10986/31009/132035-WP-FCP-New-Forms-of-Data-Processing.pdf?sequence=1&isAllowed=y

World Bank Group. (2019). Credit Scoring Approaches Guidelines. Retrieved December 14, 2021, from https://thedocs.worldbank.org/en/doc/935891585869698451-0130022020/original/CREDITSCORINGAPPROACHESGUIDELINESFINALWEB.pdf

World Bank and CGAP. (2018). Data Protection and Privacy for Alternative Data. Retrieved May 17, 2022, from https://www.gpfi.org/sites/gpfi/files/documents/Data_Protection_and_Privacy_for_Alternative_Data_WBG.pdf

Yusof Ishak Institute. (2021). The Prospects and Dangers of Algorithmic Credit Scoring in Vietnam: Regulating a Legal Blindspot. Retrieved January 12, 2022, from https://think-asia.org/bitstream/handle/11540/13169/ISEASEWP2021-1Lainez.pdf?sequence=1

Abstract

# A Checklist to Improve the Fairness in AI Financial Service: Focused on the AI-based Credit Scoring Service

HaYeong Kim* · JeongYun Heo** · Hochang Kwon***

With the spread of Artificial Intelligence (AI), various AI-based services are expanding in the financial sector such as service recommendation, automated customer response, fraud detection system(FDS), credit scoring services, etc. At the same time, problems related to reliability and unexpected social controversy are also occurring due to the nature of data-based machine learning. The need Based on this background, this study aimed to contribute to improving trust in AI-based financial services by proposing a checklist to secure fairness in AI-based credit scoring services which directly affects consumers' financial life. Among the key elements of trustworthy AI like transparency, safety, accountability, and fairness, fairness was selected as the subject of the study so that everyone could enjoy the benefits of automated algorithms from the perspective of inclusive finance without social discrimination. We divided the entire fairness related operation process into three areas like data, algorithms, and user areas through literature research. For each area, we constructed four detailed considerations for evaluation resulting in 12 checklists. The relative importance and priority of the categories were evaluated through the analytic hierarchy process (AHP). We use three different groups: financial field workers, artificial intelligence field workers, and general users which represent entire financial stakeholders. According to the importance of each stakeholder, three groups were classified and analyzed, and from a practical perspective, specific checks such as feasibility verification for using learning data and non-financial information and monitoring new inflow data were identified. Moreover, financial consumers in general were found to be highly considerate of the accuracy of result analysis and bias checks. We expect this result could contribute

* AHLab, Dept. of Smart Experience Design, Graduate School of Techno Design, Kookmin University
** Corresponding author: JeongYun Heo
AHLab, Dept. of Smart Experience Design, Graduate School of Techno Design, Kookmin University
77 Jeongneung-ro, Seongbuk-gu, Seoul, Republic of Korea
Tel: +82-2-910-5824, E-mail: yuniheo@kookmin.ac.kr
*** Trans Media Institute, Sungkyunkwan University

to the design and operation of fair AI-based financial services.

**Key Words** : Artificial Intelligence(AI), Trustworthy AI, AI-based Financial Service, AI-based Credit Scoring Service, Analytic Hierarchy Process(AHP)

Received : August 26, 2022 Revised : September 24, 2022 Accepted : September 26, 2022
Corresponding Author : JeongYun Heo

# 저 자 소 개

**김 하 영**

현재 국민대학교 테크노디자인전문대학원 석사과정에 재학 중이다. 한동대학교 ICT창업학부와 콘텐츠융합디자인학부에서 학사학위를 취득하였고, 주요 관심분야는 Human Computer Interaction, User Experience Design, Design Thinking, AI-based service 등이다.

**허 정 윤**

현재 국민대학교 테크노디자인전문대학원에서 스마트경험학과 교수로 재직 중이다. 인지적 관점에서 인간 중심의 지능형 시스템과의 상호 작용을 연구분야로 하는 증강휴먼랩을 운영하고 있다. 주요 관심 분야는 디자이너 관점에서의 인간중심의 인공지능(AI)활용 지능형 시스템 설계, 인지적 관점을 고려한 메타버스 환경에서의 사용자 경험 디자인 요소, 디자인 씽킹 관점에서의 규제 디자인 등이 있다.

**권 호 창**

성균관대학교 트랜스미디어연구소 연구교수다. 한국예술종합학교 영상이론과를 졸업하고 한국과학기술원 지식서비스공학대학원에서 박사학위를 받았다. 전산적 서사 모델링, 트랜스미디어 스토리텔링, 서사 창작 지원 시스템 등에 관한 여러 편의 논문을 썼다.