日本総研
The Japan Research Institute, Limited

AI公平性・説明可能AI（XAI）の概説と動向

2022年12月9日

株式会社日本総合研究所
先端技術ラボ

＜本件に関するお問い合わせ＞
近藤浩史（kondo.hirofumi@jri.co.jp） **間瀬英之**（mase.hideyuki@jri.co.jp） **大沼俊輔**（onuma.shunsuke@jri.co.jp）

本資料は、作成日時点で弊社が一般に信頼出来ると思われる資料に基づいて作成されたものですが、情報の正確性・完全性を保証するものではありません。また、情報の内容は、経済情勢等の変化により変更されることがあります。本資料の情報に基づき起因してご閲覧者様及び第三者に損害が発生したとしても執筆者、執筆にあたっての取材先及び弊社は一切責任を負わないものとします。本資料は、テーマとして公平性を扱っており、読者が不快に感じる場合もある説明例が用いられている箇所があります。本資料の著作権は株式会社日本総合研究所に帰属します。

© 2022 The Japan Research Institute, Limited

全体目次

© 2022 The Japan Research Institute, Limited

エグゼクティブサマリ

1. AIは性能の高さに注目が集まりがちであるが、AIをシステムに実装して利活用を促進するには、公平性、透明性、セキュリティ・堅牢性、安全性といった様々な課題や懸念事項を克服する必要がある。中でも、**「公平性」、「透明性（説明可能性など）」は多くのAI倫理原則・ガイドラインで言及されていることに加え、学術・産業界においても関連する研究が増えており、AI社会実装において肝となる論点**である。
   本レポートでは、**AI公平性、説明可能AI（XAI）について概要と動向をまとめ、技術の展望および企業に向けた推奨事項を考察・提言**した。
   特に、世の中で規制が厳しく適用される金融業界を具体例に挙げ解説している。

2. **近年、AIシステムは社会的に重要な意思決定（人事採用、犯罪予測など）に利用**される一方、AIが訓練データに含まれる人間の偏見を学習してしまうなど、「公平性」が問題になった事例が報告されている。AI公平性の研究は、**判別問題におけるグループ間の公平性を実現するための技術を中心に進展**し、今後は個人別の公平性などの、より難しい公平性を達成するための研究開発が予想される。
   また、XAIの観点では、AIとそのユーザである人間が信頼し、協調できるようにするため、AIモデルを説明・理解するための様々な手法が研究されている。提案されている手法の中から、**ステークホルダーの理解したい観点に応じて、適切な手法の使い分け**が肝要である。

3. **AI公平性、XAIの実現に向けた具体的な取り組みが始まっている。**例えば、顔認識モデルは、人種によって顔認識の精度が異なることが指摘されており、人種差別による警察の誤認逮捕などの可能性がある。そのため、例えば、Meta社は年齢・性別・見た目の肌の色・照度などがアノテーションされたデータセットを公開している。また、規制業種の金融分野では、Zest AI社が信用スコアリングに寄与する主要な要素を特定し、なぜローン審査が落ちたと判断されたのかを説明できるモデルを開発している。

4. AI公平性・XAIともに一部は実用化の段階に移っている。しかしながら、**現時点でAIを社会実装する上で求められる全てのニーズを満たしておらず、今後も基礎研究～実用化の全フェーズで研究開発が進展していく**と考えられる。
   技術動向、市場・活用動向を踏まえて、企業に向けた推奨事項として、**①新技術・ツールの調査・研究開発・技術評価、②AI倫理・ガイドラインの実行体制の整備、③システム企画段階からの検討着手、④公平性に関する社会的な合意形成、⑤適切なXAI手法・アプローチの見極め**を提言した。

© 2022 The Japan Research Institute, Limited

1.背景・導入

© 2022 The Japan Research Institute, Limited

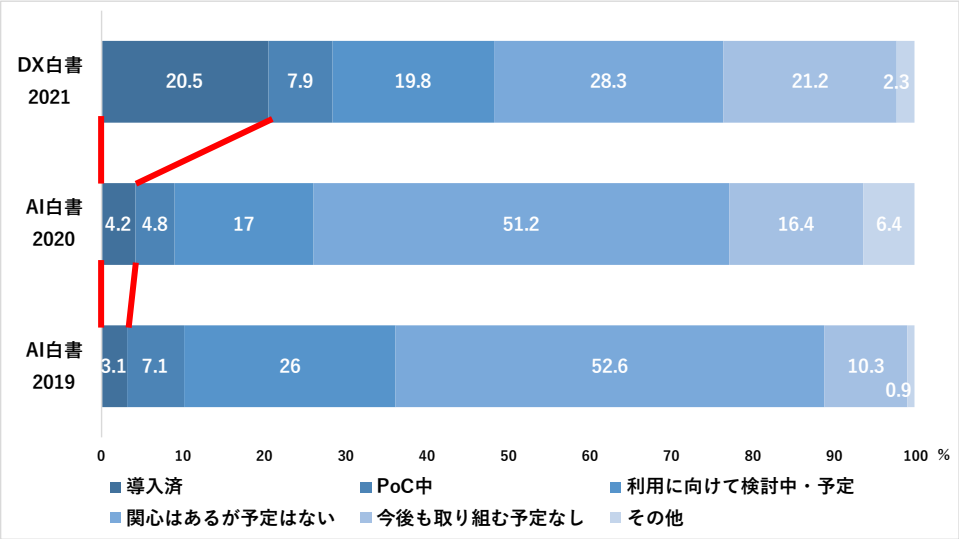
1.1 AI社会実装の進展と課題

- 近年、AI技術は急速に進展。企業のAI導入率は着実に増加しており、AI社会実装が進んでいる。
- AIは、その性能に注目が集まりがちであるが、システムに実装して利活用を促進するには、公平性や透明性、セキュリティ・堅牢性、安全性などの様々な課題、懸念事項を克服する必要がある。

AI社会実装の進展

- 米IBMの調査*1によれば、世界のAI導入率は35%に達し、加えて、42%がAIの導入を検討している。
- またDX白書2021(IPA発行)*2によれば、日本企業のAI導入率は着実に増加しており、20.5%に達する(下図)。

日本のAIの利活用状況(経年比較)



IPA「DX白書2021」 図表42-48を基に作成

*1 IBM ニュースリリース「 IBM、「世界のAI導入状況 2022年(日本語版)」を発表」(2022/7/12)
*2 独立行政法人情報処理推進機構「DX白書2021」(2021/10/11)

AI社会実装における課題・留意点

課題	概要
公平性	• 学習データから統計的に学習する性質から、AIの挙動が何らかの意味で公平でないことがある
透明性	• 利用者の安心や信頼のため、システムのデザインや運用面も含めて、第三者が理解できるようにする必要がある • システム/サービスの入出力の検証可能性及び判断結果の「説明可能性」(説明可能AI)に留意する必要がある
セキュリティ 堅牢性	• データをもとに稼働する性質から、不正なデータに対してAIが望ましくない挙動を起こすことがある • 学習データに個人情報等が含まれることがあるため、適切な管理が必要である
安全性	• AIの望ましくない挙動によって利用者や第三者に悪影響を及ぼす可能性がある • AIの性能を認識し、リスクを考慮した運用が必要である
プライバシー	• 学習データに個人情報等が含まれることがあり、プライバシーに配慮した適切な扱いが必要である
AI間の連携による影響	• AIシステム間の連携により、リスクが増幅される可能性がある
悪用の可能性	• ディープフェイク等のAIシステムの悪用を防ぐ方法を検討する必要がある

総務省「AI利活用ガイドライン」、産総研「機械学習品質マネジメントガイドライン」等を参考に作成

© 2022 The Japan Research Institute, Limited

（参考）AIの課題｜公平性、透明性

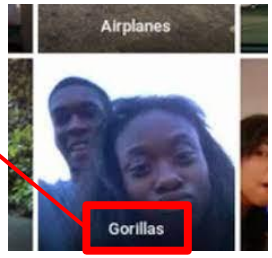
- 公平性： 学習データの偏り（バイアス）などに起因した、AIの不公平な挙動を抑止する必要がある。
- 透明性： 利用者の安心・信頼のために、システムのデザインや運用等について第三者が理解できるようにする必要がある。

公平性

- AIは基本的に、世の中に存在する実データから学習される。それらのデータには、現実に起きた歴史的・社会的な差別の結果などのバイアスを含む場合がある。学習データ等のバイアスに起因したAIモデルが示す不公平な挙動（出力）を抑止する必要がある。

Google 黒人を誤って「ゴリラ」とタグ付け

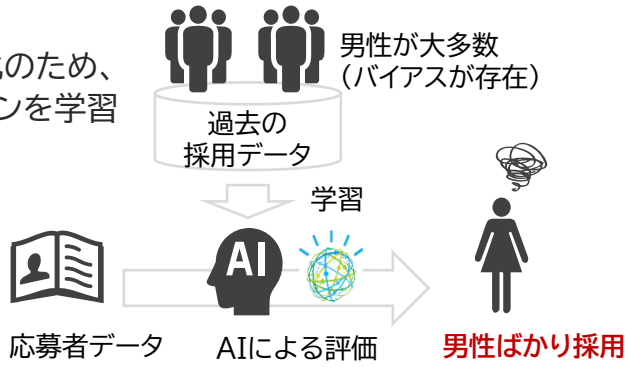
- Googleアプリは画像認識を使って自動的にタグを付ける機能を備えているが、黒人を誤って「ゴリラ」とタグ付けし、問題に。
- 学習データに偏りと偏見があったのが原因（白人の画像データに比べて、黒人のデータが少ないなど）。



Twitter @jackyalcine

Amazon AI採用の打ち切り

- エンジニア採用業務の効率化のため、過去10年分の履歴書パターンを学習し、AI採用システムを開発。
- しかし、過去エンジニア職の応募は男性ばかりだったので、女性の応募者の評価を下げる傾向に。



透明性

- 透明性は、システムのデザインや運用面、また、AIのデータ、アルゴリズムなどが第三者に理解できるようにすること。
- 各規制案やガイドラインでは、AIシステムについて責任のある判断や対応を行うため、AIがどのように動作しているのか（説明可能性）、また開発/運用の体制等について、透明性が重視されている。

アメリカ陸軍の敵、味方の戦車を識別するAI（説明可能性の例）

- AIは高精度である一方、評価や判断根拠を明確に説明できないことが多い（ブラックボックス問題）。
- この事例ではテストにおいては高い精度をあげていたが、本番に導入したところ非常に精度が低かった。
- 調査により、訓練データにバラツキがあったことが判明。
  AIは、戦車ではなく、こうした空の様子を見て判断していた。


Freitas,A.A.「Comprehensible Classification Models A position paper」

© 2022 The Japan Research Institute, Limited

1.2 AI倫理原則の議論活発化

- AIを用いたサービスやシステムを適切に管理するためにも、2016年以降、AI開発者や利用者が守るべき原則が世界各国で議論されてきた。
- 世界中のAI倫理原則に共通して重視される項目として、透明性、正義・公正、無害、責任、プライバシーが挙げられる。今後、これらを優先的に、ガバナンス・実践に向けた議論・規制化等が進むと予想される。

AI倫理原則の議論活発化

- 2016年以降、AI社会実装の本格化に向けて、国際的にAI倫理原則の議論がなされてきた。
- FRA(欧州基本権機関)の調査*1によると、2016～20年までのAI関連の政策イニシアチブ数は約350に及ぶ。我が国では、2019年に内閣府が公開した「人間中心のAI社会原則」にて、AIの社会実装にあたり留意すべき7つの「AI社会原則」が示されている。
- 欧州評議会CAHAIの調査*2によれば、世界中のAI倫理原則に共通して重視される項目として、「透明性(説明可能性など)」、「正義・公正」、「無害」、「責任」、「プライバシー」が挙げられる(右表)。
- 現状、AI倫理原則に対する明確な実践方法はないが、IT全般に係る課題(セキュリティ・プライバシー等)やAI固有のリスク(透明性、公平性等)を優先的に、今後、ガバナンス・実践に向けた議論・規制化等が進むと予想される。

*1 AI policy initiatives (2016-2020)
*2 CAHAI(2020)07-fin EN report Ienca-Vayena

(参考)AI倫理原則の分析

- 欧州評議会「CAHAI」が欧州内外116個のAI原則の文書を調査。

倫理原則	該当数	含まれるコード
Transparency (透明性)	101	Transparency, explainability, interpretability, communication
Justice and fairness (正義と公平)	97	Justice, fairness, consistency, inclusion, equality, equity, (non-)bias, (non-)discrimination, diversity
Non-maleficence (無害、安全性)	84	Non-maleficence, security, safety, harm, protection
Responsibility (責任)	79	Responsibility, accountability
Privacy	74	Privacy, confidentially
Beneficence	58	Benefits, peace
Freedom and autonomy	48	Freedom, autonomy, consent, choice
Trustworthiness	41	Trust, trustworthiness
Sustainability	20	Sustainability, nature
Dignity	20	Dignity
Solidarity	10	Solidarity, social security

欧州評議会 CAHAI「Ethics Guidelines for Trustworthy AI」を基に作成

© 2022 The Japan Research Institute, Limited

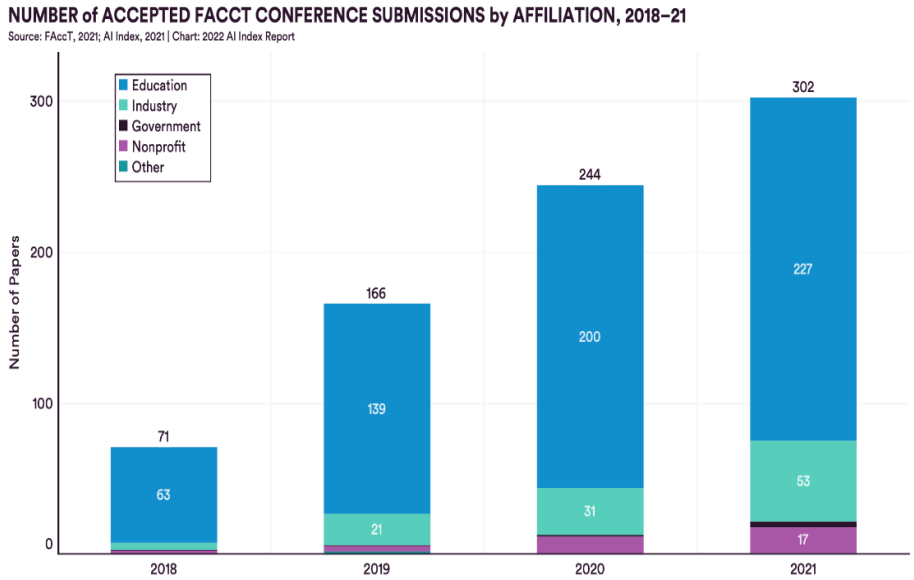
1.3 本レポートで採り上げる技術トピックと構成

- AIに関する原則・ガイドラインの多くが「公平性」「透明性（説明可能性など）」に言及している（前頁）ほか、学術・産業界においても関連する研究が拡大している（下図）。また、AIガバナンスを強化するツールとして活用可能な製品・サービスも登場し始めている。
- 本レポートではAIの倫理原則において肝となる、「公平性」「説明可能AI（以降、XAI*1という）」に焦点をあて、技術動向[第2章]、市場動向[第3章]、今後の展望[第4章]の構成で解説する。特に、世の中で規制が厳しく適用される金融業界を具体例に挙げ解説している。

*1 Explainable AI の略

公平性/説明責任/透明性に関する論文の件数

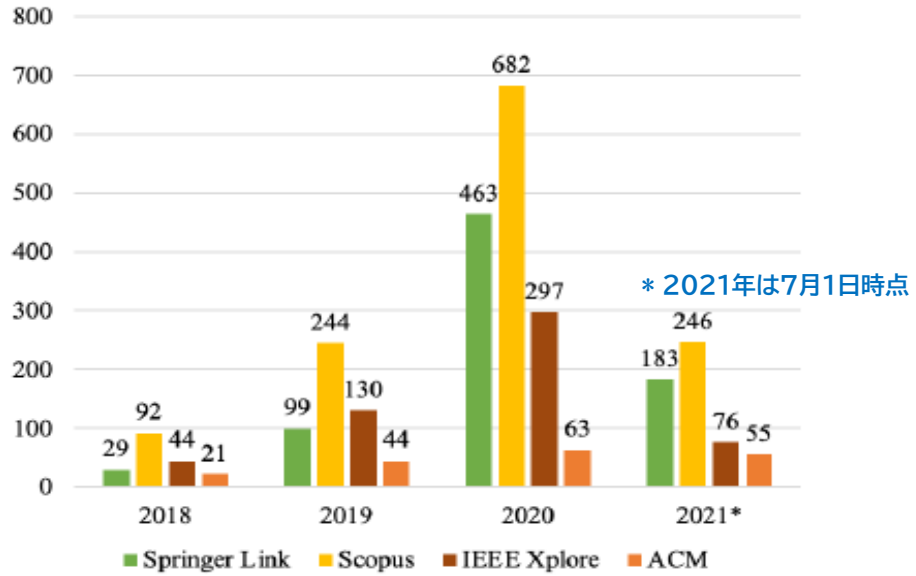
- 公平性・説明責任・透明性に関する国際会議（ACM FAccT）に採録された論文の推移。



Artificial Intelligence Index Report 2022, https://aiindex.stanford.edu/wp-content/uploads/2022/03/2022-AI-Index-Report_Master.pdf のFigure3.3.1を抜粋

説明可能性に関する学術雑誌の件数

- 説明可能性に関する学術雑誌（Springer Link、Scopus、IEEE Xplore、ACM）の件数推移。



＊2021年は7月1日時点

Mir Riyanul Islam et al, A systematic Review of Explainable Artificial Intelligence in Terms of Different Application Domains and Tasks, Applied Science, Jan 2022 のFigure1 (b) を抜粋

© 2022 The Japan Research Institute, Limited

2.技術動向・解説

© 2022 The Japan Research Institute, Limited

2.1.1 AIにおける公平性とは

- 近年、AIシステムが社会的に重要な意思決定（人事採用、犯罪予測など）に使用される事例が増加。一方で、AIが訓練データに含まれる人間の偏見を学習してしまうなど、公平性が問題になった事例も多数報告されている。
- AIシステムが社会で安心して使用されるには、システム特性などに応じた公平性基準を設定し、その基準を満たすようにAIモデルを学習・運用するといった、AIシステムの公平性実現が必要となる。

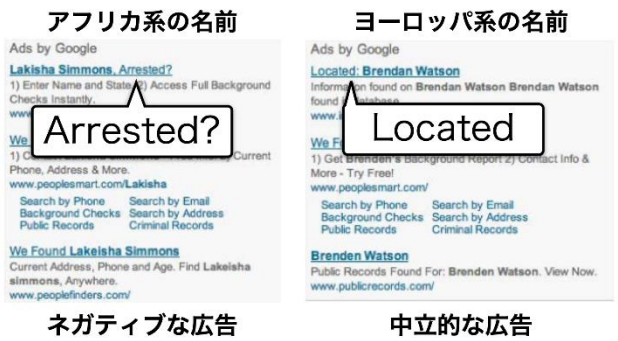
公平性が問題になった事例

- 米国の再犯リスク予測において、人種ごとに判別精度は同等でも、アフリカ系の誤検知率が高い（見逃し率は低い）。

	白人	アフリカ系
高リスクと判定されたが再犯せず（誤検知）	23.5%	44.9%
低リスクと判定されたが再犯した（見逃し）	47.7%	28.0%

J. Angwin, J. Larson, S. Mattu, and L. Kirchner. Machine bias, 2016.
https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

- 検索エンジンGoogleで人名を検索した際に表示される広告を集計したところ、アフリカ系の名前の場合、ネガティブな広告が表示される傾向があった*1。



AIにおける公平性の扱う範囲

- AIシステムが実現すべき公平性は、社会・文化的背景などに依存し、曖昧かつ多様なニーズがあるため、明確に定義することは困難である。

【例】採用の場合、例えば以下のような公平の考え方がありうる
① 採用される男女の数が等しくなること
② 応募者の中での採用率が男女で等しくなること
③ 応募の機会が男女間で等しく与えられていること

- 現状のAI公平性の議論では、曖昧かつ多様な公平性のニーズの中から、「公平であること」を数式化可能な形で定義し、その定義を満たす公平なAIシステムの開発を目指す。また、その定義を満たすようにAIシステムを運用することも求められる。
- なお、公平性に関する社会的な要請に明確なものはなく、今後、AI公平性の議論が深まり、社会に浸透していく過程で、合意が得られていくと期待されている。

公平性に配慮した学習とその理論的課題
https://ibisml.org/ibis2018/files/2018/11/fukuchi.pdf

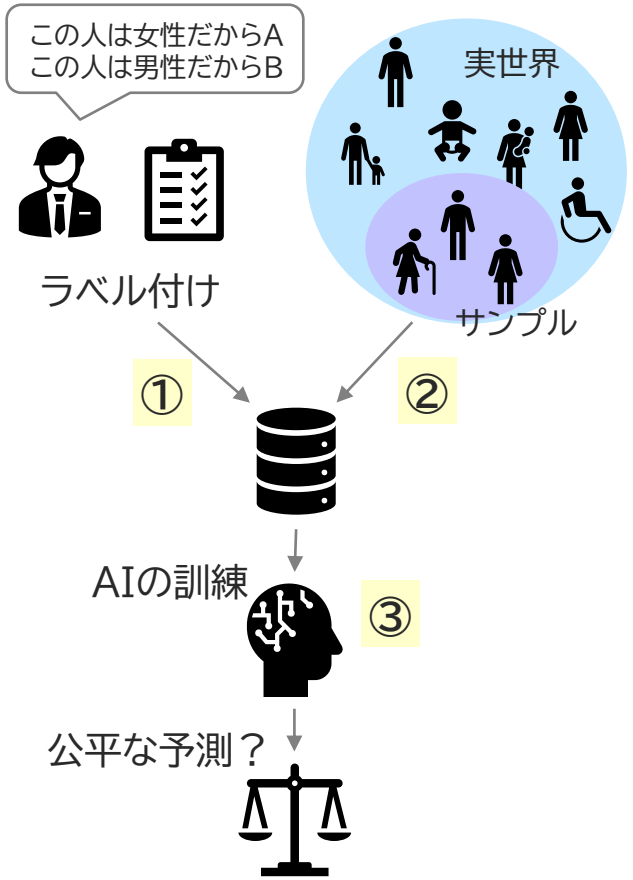
*1 Sweeney, L.: Discrimination in Online Ad Delivery, *Communications of the ACM*, Vol. 56, No. 5, pp. 44–54 (2013)

© 2022 The Japan Research Institute, Limited

（参考）AIの不公平な判断の発生原因

- AIが不公平な判断をする原因は、AIモデルを学習する際に取り込んでしまうバイアスの存在である。
- 代表的なバイアスとして、「データ自体のバイアス」「サンプル選択バイアス」「帰納バイアス」の3種類が存在する。

バイアスが発生する場所とバイアスの種類



バイアスの種類	概要
① データ自体のバイアス（Data bias）	訓練データをラベル付けした人の偏見などにより、訓練データの特徴に偏りが生じる。 【例】 • 過去のローン査定結果を学習させると、過去の意思決定者のバイアスを反映したモデルになる。
② サンプル選択バイアス（Sampling bias）	訓練データが予測対象の全体を十分に反映していないために、偏りが生じる。 【例】 • ローンの審査モデルを作る場合、過去のローン審査落ちした顧客のデータは削除されている場合があるため、正しいモデルが作れない。
③ 帰納バイアス（Inductive bias）	機械学習アルゴリズムが汎化のために採用している仮定が、現実世界の現象と外れているために偏りが生じる。 【例】 • 少数派のローン完済データは少数ゆえに学習において無視されやすい • 多くの機械学習手法はテスト精度を高めるためにシンプルなモデル（Ridge正則化、Lassoなど）を学習しやすい。元々そうである以上に少数者のローン完済確率を低く見積もりやすい。

© 2022 The Japan Research Institute, Limited

2.1.2 AI公平性の導入に向けて

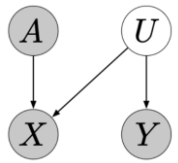
- AI公平性の実現に向けては、AIシステムの開発段階に応じて検討すべき事項がある。
  - 企画/設計 ・・・ 実現すべき公平な状態を定義し、それらを測定するための指標を定める
  - 開発 ・・・ 訓練に使用するデータの公平性、モデルの公平性を担保するための施策を適用する
  - 運用/モニタリング ・・・ 定期的なモニタリングと不芳時に対処する

フェーズ	検討項目
企画/設計	• 過去の公平性に関するインシデント調査など通して、提供するシステムが達成すべき公平性の要件を分析し、公平な状態を定義する(→ P.13-14解説)。 • 公平性の定義に基づいて、公平な状態を計測するための指標を定める。
開発	• AIモデルの訓練データについて、差別を維持・助長するようなデータセットになっていないかを確認する。 • 定めた指標値をもとに、公平な状態が達成されるよう、AIモデルの訓練データ、モデルの学習アルゴリズム、モデル出力などに対して何らかの介入を検討する(→ P.15-18解説)。
運用/モニタリング	• 定めた指標値をもとに、公平性が達成できているかをモデルの実運用において定期的にモニタリングする仕組みを整備する。 • 閾値を超える変化があった場合はアラートを発出するなどし、対処方法を検討できる仕組みを構築する。

© 2022 The Japan Research Institute, Limited

2.1.3 公平性の定義の体系

- 提案されている代表的な公平性の定義と、公平性の度合い測定する指標例を以下に整理。
  なお、以降、配慮が必要な属性（例：人種、性別）のことを「センシティブ属性」と呼ぶ。
- センシティブ属性値の異なる集団の間で差別しないことを目的とした定義を「グループ間の公平性」、個人・個別の要因に対処して差別しないことを目的とした定義を「個別的公平性」と分類する。

分類		公平状態定義	説明	公平状態が満たされない例	指標例
無知による公平性		手続きに基づく公平性	センシティブ属性を特徴量として使わない*1	履歴書に出身や写真を載せ、それらに基づいて採否判断をする	手続きに基づくため無し
グループ間の公平性	データバイアス	結果公平性（統計的パリティ）	センシティブ属性間で予測ラベルの割合が同じ	就職希望者のうち、採用するべきと予測する割合が性別で異なる	Statistical Parity Difference/ Disparate Impact
	帰納バイアス	等価オッズ、分離、機会均等	真のラベルが同じなら、予測の誤検知率/見逃し率等がセンシティブ属性間で同じ	出所後の再犯予測において、白人は見逃し率が高く、黒人は誤検知率が高い*2	真陽性率差/ 偽陽性率差
		充足性	予測ラベルが同じならその予測の正解率がセンシティブ属性間で同じ	保険の不正検知において、偽発見率（不正と判定され調査したが不正ではなかった割合）が人種で異なる	偽発見率差/ False Omission Rate Difference
個別的公平性		個人公平性（シチュエーションテスト）	合理的特徴（人種や性別を除く、それに基づく異なった扱いが合理化される特徴）が似ていれば予測も似ている	クレジットスコアが同等の男女間で、女性の方が与信枠を少なく設定される*3	一貫性/ 一般化エントロピー指数
		因果に基づく公平性	センシティブ属性がもし異なっていたとした場合に、他変数への因果効果も加味した上で予測に影響がないこと。ただし因果的に独立した変数（隠れた背景因子を含む）は固定した状態とする。	（1）危険選好因子（U）を持つ人は、赤い車（X）を好み、かつ事故率（Y）が高いとする。（2）ある人種（A）は赤い車（X）を好むとする。この場合、単に「赤い車は保険料を高く」設定する仕組みでは、人種（A）が事故率（Y）の原因ではないにも関わらず、人種Aは赤い車を買いやすいために保険料が高く設定される	（研究途上につき、代表的な指標なし） 

*1 なお、センシティブ属性だけを直接取り除いても、センシティブ属性と相関が高いデータが残っていた場合、間接的にセンシティブ属性の情報をAIが学習してしまうことがある（いわゆるレッドライン効果）
*2 Angwin, J., Larson, J., Mattu, S., and Kirchner, L. Machine Bias. www.propublica.org/article/machine-biasrisk-assessments-in-criminal-sentencing, May 2016.
*3 Vigdor, N. Apple Card Investigated After Gender Discrimination Complaints. The New York Times, November 2019. ISSN 0362-4331.

© 2022 The Japan Research Institute, Limited

2.1.4 公平性定義の使い分けに関する整理

- AIシステムのユースケースに応じて適切な公平性定義を採用し、使い分ける必要がある。
- まずは、世の中で研究が進むグループ間の公平性を検討する(Step1)。それでも不十分な場合は、個人公平性や因果に基づく公平性を検討する(Step2,3)。

検討段階	条件	採用を検討する公平性定義
Step1	• センシティブ属性に関するグループ単位で大まかに公平性を実現する場合。 その中でも大別して2種類の使い分けが存在。 1. 扱うデータに信頼できるラベルが存在しない場合。 【例】採用面接での採否ラベルに男女間で偏りがあり、その後の営業成績などの合理的な結果データが存在しないが、採否を判定するモデルを作りたい。 2. 扱うデータに信頼できるラベルが存在するが、データの傾向を学習によって助長することを避けたい。 【例】再犯予測で、その後の再犯データが存在するが、機械学習によって人種ごとに偏った出力となることを防ぎたい。	1. 公平性定義として「統計的パリティ」など結果公平性(センシティブ属性と予測の独立性に関する指標)の採用を検討する。ただし、センシティブ属性によって営業成績などに真に違いがある場合には逆差別となるリスクがある。 2. 「等価オッズ」など帰納バイアスを防止する定義の採用を検討する。
Step2	• グループ全体での平均的な公平性からより一歩進んで、より個々人の結果を公平にすべき場合。 【例】男性は女性と比べてスコア(合理的説明変数)のばらつきが大きく、グループ公平性を課すと同程度のスコアでも高得点帯では男性不利、低得点帯では女性不利となる。	個人ごとの公平性を検討。
Step3	• 同じ説明変数の値でも、その背後にセンシティブ属性の因果的な影響が想定され、扱いを分けるべき場合。明確な因果関係の想定に基づいて間接的な影響と、合理的に説明可能な影響とを分離したい場合。 【例】赤い車は危険運転をしやすい人によって所有されることが多いが、民族的な理由での所有もあり得る。	因果に基づく公平性を検討。

(参考)統一された定義ではなく個人ごとに異なる公平性への要望がある場合への対処など、さらに発展的な議論も存在。
ただし因果に基づく公平性などは、定義も含め、研究途上という側面が強いため現状では慎重な検討を要する

© 2022 The Japan Research Institute, Limited

2.1.5 手法の分類

- 定義した公平性を実現するための手法は複数存在する。特に研究が多いのは、判別タスクにおける公平なモデルの学習手法である。
- 次頁から、判別タスクにおけるグループ間の公平性に着目した公平なモデルの学習手法について解説する。特に、処理段階の分類(前処理、学習中、後処理)に基づき、それぞれ解説する。

公平性に配慮した機械学習手法の分類

- 手法は①不公平を発見するか、②不公平を防止するか、という2つの軸に大分類できる。
- それらが更にいくつかの軸で分類できるが、研究が多いものは判別タスクによる公平なモデルの学習である。

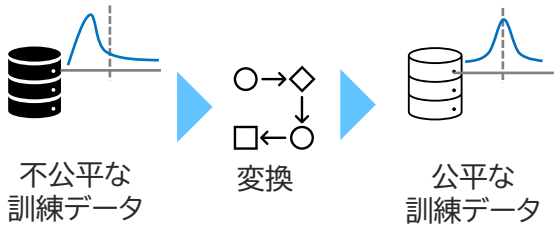
大分類	小分類	説明
不公平の発見	データセットの不公平の検知	データセット中で不公平となっているグループやデータを検知する。
	モデルの不公平の検知	モデルから不公平な出力結果を検知する。
不公平の防止(公平なモデルの学習)	処理段階での分類	モデル学習の処理段階に着目した分類。前処理(pre-process)/学習中(in-process)/後処理(post-process)の3つに大別。
	タスクでの分類	構築するモデルが取り組むタスクによる分類。判別 / 回帰 / クラスタリング(教師なし学習) / 画像・テキスト生成 / 推薦など。

Kamishima, T. “Fairness-aware machine learning and data mining.” https://www.kamishima.net/archive/faml.pdf をもとに作成

処理段階での分類イメージ

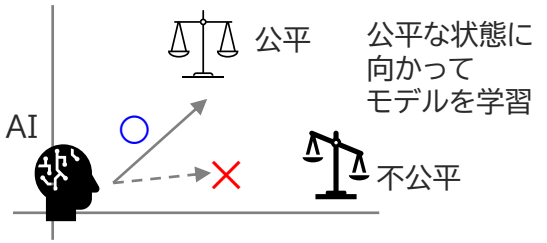
前処理

モデルを学習する前に、センシティブ属性間で不公平がなくなるように訓練データを編集
(→ P.17で解説)



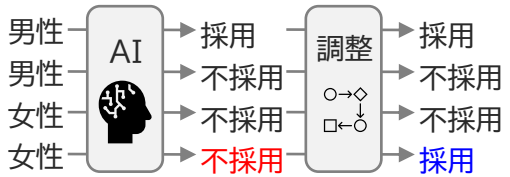
学習中

モデルの学習中に不公平をなくすような制約をかけて、モデルを学習
(→ P.16で解説)



後処理

訓練済みモデルの出力をセンシティブ属性間で不公平がなくなるように調整
(→ P.17で解説)



公平性を実現するように出力を後から調整

© 2022 The Japan Research Institute, Limited

2.1.6 公平な判別モデルの学習(処理段階:学習中)

- 学習中に公平性を考慮する手法は、一般に性能が高く、精度と公平性を高いレベルで両立することができ、かつ公平性の種類に対しても汎用性が高い。
- 一方、基本的にはモデルの種類(訓練アルゴリズム)ごとに専用に実装が必要である。訓練済みモデルのみを用いて調整できない(モデルをゼロから訓練する場合にのみ適用可能)など、簡便さには欠ける。

制約ベースの手法

罰則ベース(Prejudice Remover[*1])

- 学習時に誤差に加えて公平性指標との重み付き和を最適化。
- モデルごとに実装が必要。 $\min_f \mathrm{Err}(f) + \eta\,\mathrm{Unfair}(f)$

制約ベースのコスト考慮型学習への帰着(Exponentiated Gradient Reduction[*2])

- 公平性指標(複数でも可)が一定以内となる範囲で誤差最小化。

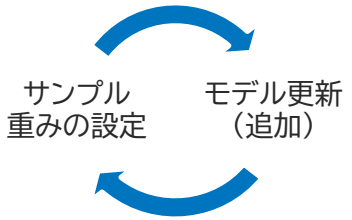
$$\min_f \mathrm{Err}(f)\ \mathrm{s.t.}\ \mathrm{Unfair}(f) \le \eta$$

- コスト考慮型学習(サンプル重み付き学習)を用いる。

1. 未充足の制約式について、対応する予測コストを高く設定

【統計的パリティの場合の例】全体の採用率が女性の採用率より少ない予測となっている場合、女性を不採用と予測するコストを高く設定

2. 任意のコスト考慮型学習器(サンプル重みを設定できる機械学習手法)を使って予測器を訓練(アンサンブルの1モデルとして追加)

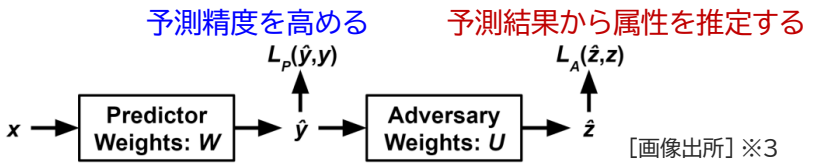


3. 収束まで上記を繰り返す

敵対的な学習による手法

敵対的バイアス除去(Adversarial Debiasing [*3])

- 予測結果からセンシティブ属性が逆推定できないようにすることで結果公平性を高める。
- 予測器(パラメタW)は予測精度を高めつつ、予測結果からセンシティブ属性zが推定されない予測値を出力するように訓練する。
- 敵対的判別器(パラメタU)は予測器の出力した予測値からセンシティブ属性zを推定するように訓練する。
- 男性の方が採用割合が大きいなど予測値の分布に差があれば敵対的判別器が予測値からセンシティブ属性を逆推定できる。裏を返せば、敵対的判別器が逆推定できないなら公平である。



[画像出所] ※3

*1 T. Kamishima, S. Akaho, H. Asoh, and J. Sakuma, "Fairness-Aware Classifier with Prejudice Remover Regularizer," Joint European Conference on Machine Learning and Knowledge Discovery in Databases, 2012.
*2 A. Agarwal, A. Beygelzimer, M. Dudik, J. Langford, and H. Wallach, "A Reductions Approach to Fair Classification," International Conference on Machine Learning, 2018.
*3 Zhang, Brian Hu, Blake Lemoine, and Margaret Mitchell. "Mitigating unwanted biases with adversarial learning." Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society. 2018.

© 2022 The Japan Research Institute, Limited

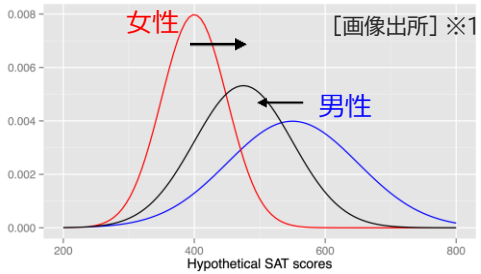
2.1.7 公平な判別モデルの学習(処理段階:前処理 / 後処理)

- 前処理法、後処理法は学習器に関係なく適用できる点で適用範囲が広い。
  - 前処理法:基本的にデータに内在する不公平性を除去するものであり、結果公平性との関連が強い
  - 後処理法:MLパイプライン全体をブラックボックスとして扱える点で汎用性が高い。

前処理における手法

Disparate Impact Remover*1

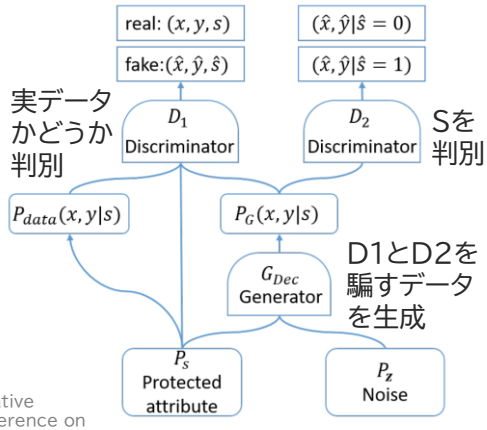
- 特徴量分布(例: テストの得点)がセンシティブ属性(例: 性別)間で類似するように編集(ただし、同性内での順位は保存)。
- これにより Disparate Impact (結果不公平性)が低減される。



*1 Feldman, Michael, et al. "Certifying and removing disparate impact." Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining (KDD). 2015.

FairGAN*2

- 実データらしさが高く、かつセンシティブ属性(S)を判別できないようなデータ (x, y) を生成する生成器Gと、実データかどうか判別する判別器D1およびSを判別する判別器D2を敵対的に訓練する。
- 訓練済みのGによって生成されたデータは(結果)公平なデータになっていると考えられる。



[画像出所] ※2

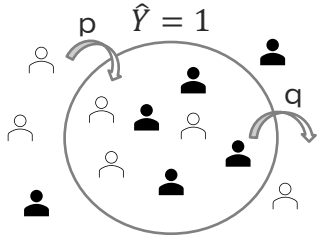
*2 Xu, Depeng, et al. "Fairgan: Fairness-aware generative adversarial networks." 2018 IEEE International Conference on Big Data (Big Data). IEEE, 2018.

後処理における手法

Equalized Odds Postprocessing*3

- 等価オッズを達成するために出力ラベルを確率的に変更する。

【例】以下の確率 (p, q) をうまく設定すると等価オッズを達成できる
- 👤 の非再犯予測者を確率pで再犯予測者に変更
- 👤 の再犯予測者を確率qで非再犯予測者に変更

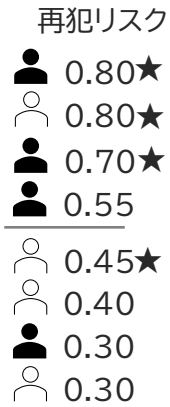


*3 M. Hardt, E. Price, and N. Srebro, "Equality of Opportunity in Supervised Learning," Conference on Neural Information Processing Systems, 2016.

Reject Option Classification*4

- ソフトな実数値のラベルを出力する予測手法において、境界付近($|Y^ - 0.5| \leq \theta$)の予測をセンシティブ属性ごとに正例側/負例側に倒す。

【再犯予測を例にしたイメージ(右図)】
右図で0.5を閾値にすると再犯予想は👤に偏る。そこで、👥は再犯リスク0.4以上を再犯予想に、👤は再犯リスク0.6以上を再犯予想にすると、★印の人物が再犯予想となり、バランスが取れる



*4 F. Kamiran, A. Karim, and X. Zhang, "Decision Theory for Discrimination-Aware Classification," IEEE International Conference on Data Mining, 2012.

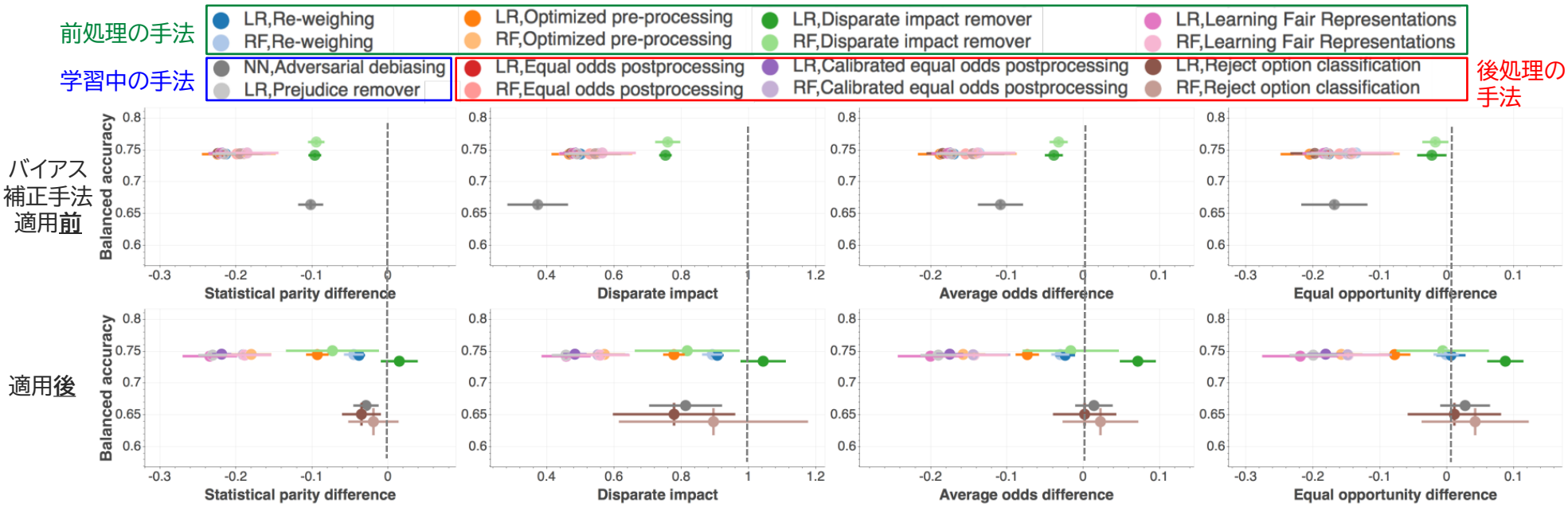
© 2022 The Japan Research Institute, Limited

2.1.8 公平な判別モデルの学習手法の性能比較

- 公平な判別モデルの学習手法ごとに、適用前後で得られる精度と公平性指標に違いが見られるため、適切な手法を選択する必要がある。
- IBM社の研究*1によれば、精度は多くの手法で変化しないが、公平性指標には違いが見られた。データセットごとによって異なるが、おおまかな傾向として、①後処理による手法は簡便である一方、しばしば精度低下を伴う、②前処理および学習中の手法は大きく性能劣化しないことが多い。詳細は論文*1を参照。

*1 Bellamy, Rachel KE, et al. "AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias." *IBM Journal of Research and Development* 63.4/5 (2019): 4-1.

性能比較例(Adult Census Incomeデータセット、センシティブ属性:人種)



[画像出所] *1

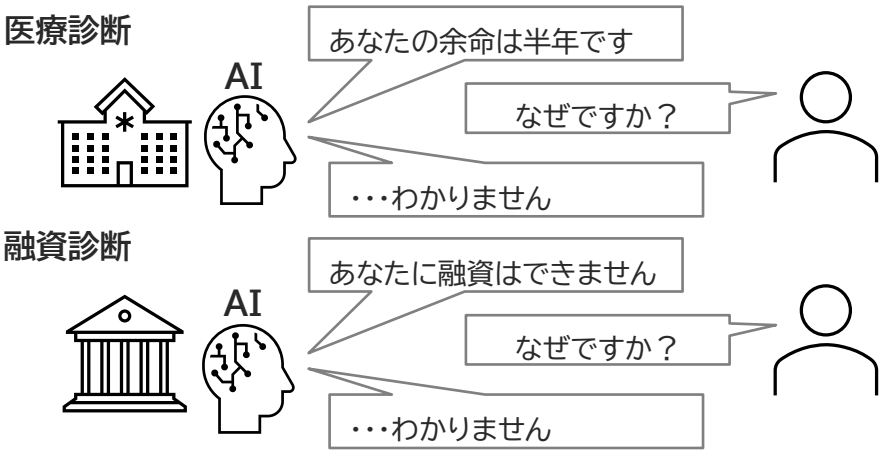
© 2022 The Japan Research Institute, Limited

2.2.1 AIにおける説明可能性とは

- AI社会実装においては、ユーザである人間がAIを信頼し協調できることが重要。特に、医療や金融などの社会的な影響が大きい領域では、XAI(説明可能AI)が必要となる。
- 規制やガイドラインの検討等も進んでおり、それらへの対処においても有用と考えられる。

XAIの必要性

- AIは高精度である一方、評価や判断根拠を明確に説明できないことが多い(ブラックボックス問題)。
- そのため、社会的インパクトが大きい領域(医療、金融など)におけるAI活用では、ユーザがAIを信頼するために、AIの挙動が理解可能であることが要求される。
- ユーザである人間が適宜、AIの出力を修正したり、対話的に操作したりする際に、出力結果を理解して行うことで、よりよい協調につながる。



規則等で求められる透明性との関係

- AIガバナンスに関する規制やガイドラインの検討、整備が進んでおり、モデルのリスク評価や、システムの運用面、データ、アルゴリズム等の透明性が求められている。
- モデルのリスク評価や挙動を理解・説明するため、XAI技術が有用と思われる。

透明性やリスク評価等に関連する規制やガイドライン例

	説明
EU GDPR	EUの一般データ保護規則(GDPR)*1では、自身の個人データに基づく意思決定がAI等によってされた場合、その決定に関連する情報提供を受ける権利が記されるなど、データや意思決定に関する、透明性(説明責任など)が重視されている。
金融庁「モデル・リスク管理に関する原則」	日本の金融庁が公表している「モデル・リスク管理に関する原則」(2021年11月)*2では、モデル・リスクを適切に管理することを目的として、実効的なけん制が行われるための体制構築を示している。モデルのリスク評価と、判断したリスクを踏まえた対応が期待される。

*1 個人情報保護委員会『一般データ保護規則(GDPR)の条文』(仮日本語訳)
https://www.ppc.go.jp/files/pdf/gdpr-provisions-ja.pdf
*2 金融庁『モデル・リスク管理に関する原則』
https://www.fsa.go.jp/common/law/ginkou/pdf 02.pdf

© 2022 The Japan Research Institute, Limited

2.2.2 取り上げるXAI技術

- XAIは、AIモデルの挙動を人間に理解しやすくするモデルや手法であり、説明対象や考え方に応じて様々なアプローチの手法が研究開発されている。
- このレポートでは、代表的な手法としてモデル自体を解釈しやすいモデルとするアプローチ（解釈可能モデル）、入出力の関係を可視化するアプローチ、ブラックボックスモデルを解釈しやすいモデルで近似するアプローチ（Surrogate Model）、仮想的なサンプルで利用者にフィードバックを与えるアプローチを取り上げる。また画像データにおける手法も取り上げる。

	手法名	説明
解釈可能モデル	線形回帰 決定木	• 線形回帰や決定木はモデル構築の方法から、人間に理解しやすい古典的な手法。
特徴量と予測値の関係の可視化	ICE, PDP	• モデルの構造は考慮せず、特定の特徴量を変動させた時の出力を観察することで、特徴量と予測値の関係性を可視化する手法。
近似モデルによる説明可能性（Surrogate Model）	Global/Local Surrogate Model	• 説明したいモデルを解釈しやすいモデルで近似することで説明する手法。 • モデル全体を近似する手法をGlobal、個別の予測ごとに近似する手法を「Local Surrogate Model」と呼ぶ。
	LIME	• LocalなSurrogate Model手法。説明対象のデータの周辺に注目して、モデルの振る舞いを近似する。
	SHAP	• LocalなSurrogate Model手法。協力ゲーム理論に基づいて、予測結果に対する個々の特徴量の寄与度合いを計算する。
仮想サンプルによる説明可能性	Counterfactual Explanations/ Adversarial Examples ※	• Counterfactual Explanationsは、実際と異なる望ましい結果をもたらすような、微小に変化したサンプルを利用して挙動を説明する手法。 • Adversarial Examplesはモデルを騙すような微小な変化を加えたサンプルのこと。
画像データにおける説明可能性の例	Saliency Map	• 画像処理モデルの予測に対する各ピクセルの重要度を可視化する。

※ Adversarial Examplesは説明可能性を目的とした手法ではないが、Counterfactualと関係があるため合わせて取り上げる

Christoph Molnar, Interpretable Machine Learning, https://christophm.github.io/interpretable-ml-book を参考に作成

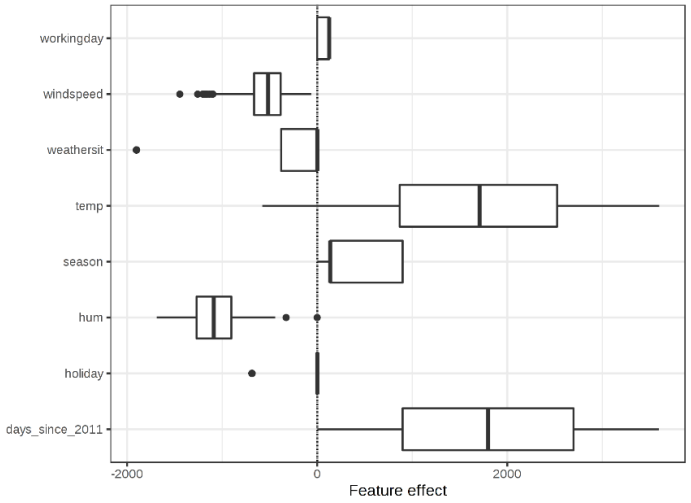
© 2022 The Japan Research Institute, Limited

2.2.3 解釈可能モデル

- モデルの計算方法から、入力特徴量と予測値の関係が解釈可能であるモデルのこと。線形回帰や決定木などが代表的である。
- 古典的な手法であるが、計算過程は人間にとって理解しやすく、想定しない挙動を起こす可能性が低い。精度よりも安定性を重要視する場面においては依然として有用と考えられる。

線形回帰

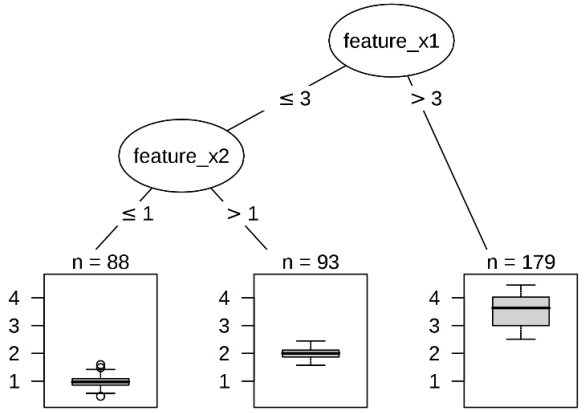
- 線形回帰では特徴量を重み付けした和から予測する。
- 特徴量の重みは、その特徴量の重要度として解釈できる。
- 各特徴量の重み付けした値は、各サンプルにおいて特徴量が予測にどの程度の影響を与えたかを説明するのに利用できる。



サンプルデータで特徴量の影響をプロットした図

決定木

- 特徴量の値によって条件分岐を行い、各データをクラスターに分割する。
  - クラスターのごとに予測値を出力する。
- この条件分岐によってモデルの挙動を説明することができる。
- 全ての分岐において、その分岐が存在しなかった場合と比べてどれくらい精度が向上したかを評価することで、それらの分岐に対応する特徴量の重要度を計算することができる。



サンプルデータで学習された決定木のイメージ図

Christoph Molnar, Interpretable Machine Learning, https://christophm.github.io/interpretable-ml-book

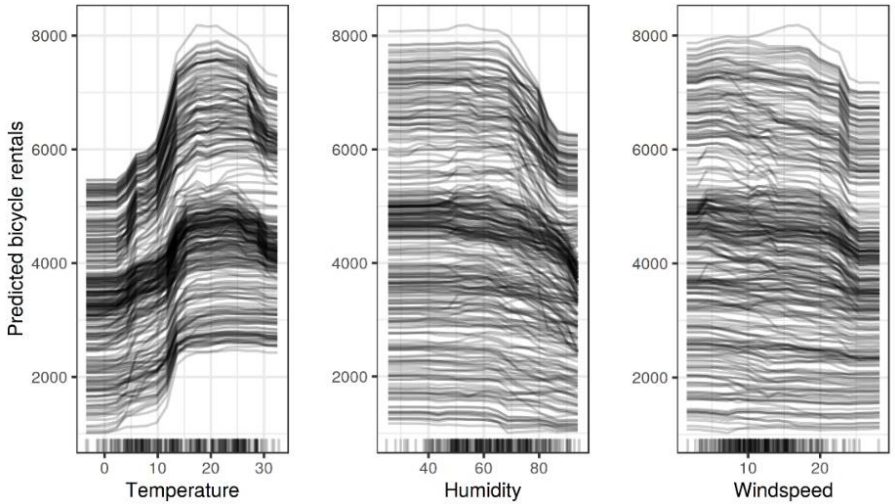
© 2022 The Japan Research Institute, Limited

2.2.4 特徴量と予測値の関係の可視化

- 特徴量の値と予測値の値の関係を可視化することで、モデルの挙動を解釈する手法。モデルの入出力関係だけに注目するため、ブラックボックスモデルでも適用できる。
- ただし、特定の特徴量の値を人工的に変動させることで、現実的ではないデータを生成してしまう場合があるため、実際とは異なる関係性が得られることがある。

Individual Conditional Expectation（ICE）

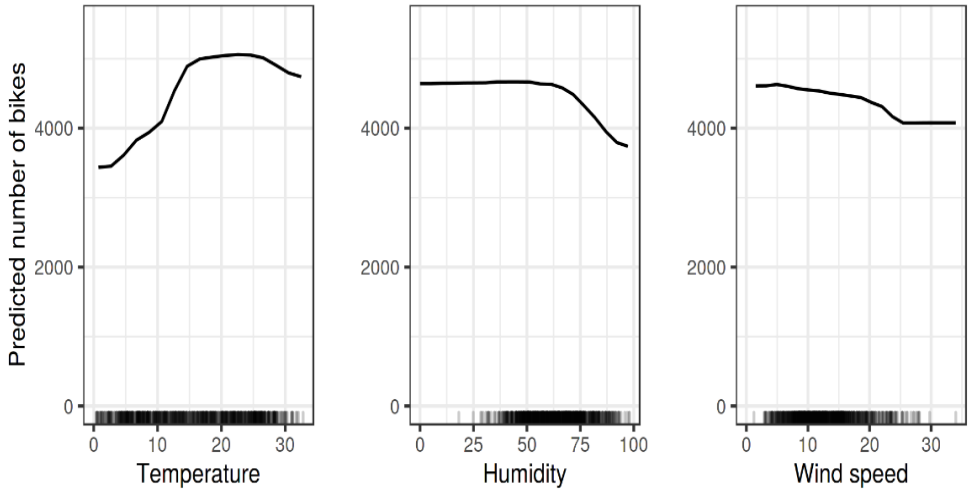
- 全てのデータにおいて、評価対象の特徴量の値のみを変動させて、モデル予測値との相関関係を可視化。（他の特徴量は固定）
- 例えば、下左図では「温度」が上昇するほど予測値が非線形的に増加していることが分かる。



ICEではサンプル数に対応して個別に曲線が出力される。
左図は「温度」の入力を人工的に変化させた時の予測値の変化。中央図は「湿度」、右図は「風速」を変化させたもの。

Partial Dependence Plot（PDP）

- 評価対象の特徴量を変動させて得られたモデル予測値を、平均化して示す。
- PDPはICEにおけるモデル予測値を平均化したもので、両者は併せて扱われることが多い。



左から「温度」「湿度」「風速」の入力を人工的に変化させた時の予測値の変化。各サンプルの結果が平均化されている。

Christoph Molnar, Interpretable Machine Learning, https://christophm.github.io/interpretable-ml-book

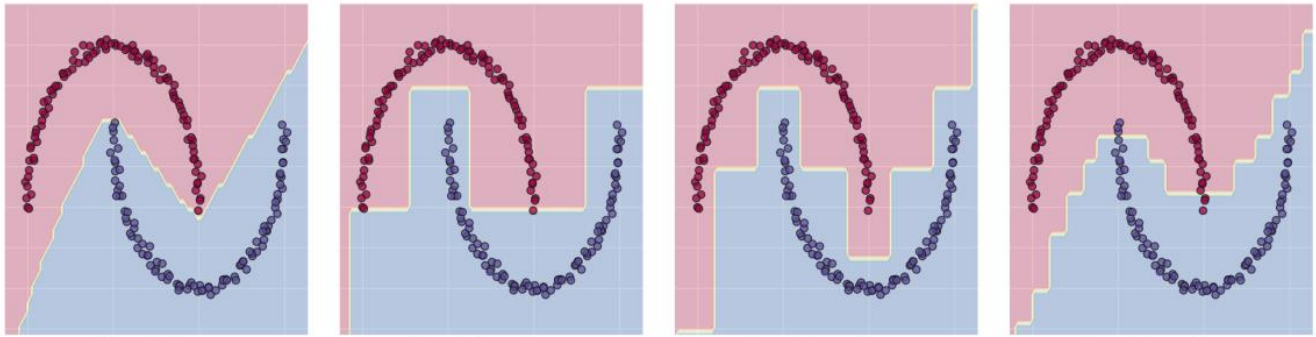
© 2022 The Japan Research Institute, Limited

2.2.5 Surrogate Model

- 説明対象のブラックボックスモデルを、線形回帰や決定木などの説明可能性の高いモデル(Surrogate Model)で近似することにより、間接的に説明する手法。モデル全体の挙動を説明する「Global Surrogate Model」と、各予測について挙動を説明する「Local Surrogate Model」がある。
- Surrogate Modelは近似性能が低い場合は十分に解釈できない点や、あくまで近似モデルであるという点に注意が必要。

Global Surrogate Model

- 説明対象のブラックボックスモデル自体の大域的な挙動を説明。説明変数とそれを入力とするブラックボックスモデルの予測値を学習データとして近似モデルを学習する。
- 下図は、二つの半円状のデータ集合からニューラルネットによるブラックボックスモデルを生成し、Surrogate Modelで近似した例。



Rafael Poyiadzi et al., Understanding surrogate explanations: the interplay between complexity, fidelity and coverage, arXiv preprint arXiv:2107.04309, 2021
説明のため赤枠、赤字、青字を追記

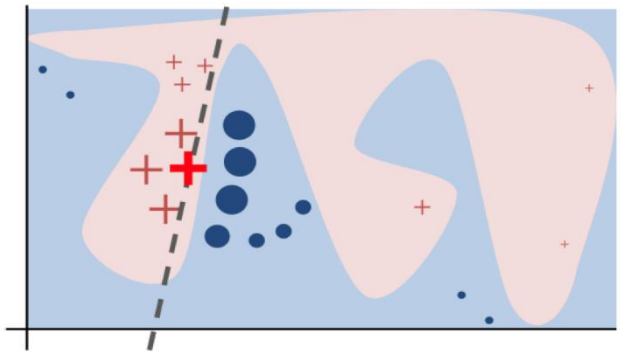
© 2022 The Japan Research Institute, Limited

2.2.6 Surrogate Model - LIME

Local Surrogate ー LIME(local interpretable model-agnostic explanations)

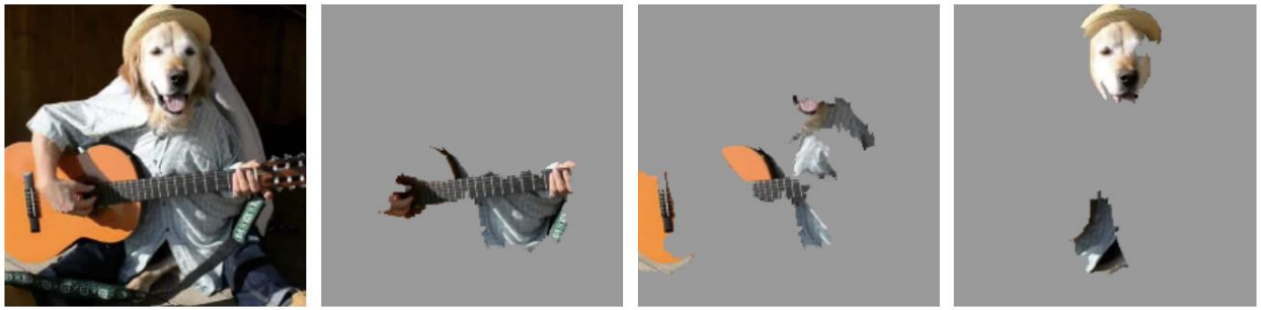
- 各々のデータ入出力の挙動を説明するために個別に作られたSurrogate Modelを「Local Surrogate Model」と呼ぶ。代表的な手法に「LIME」と「SHAP」がある。
- LIMEは、説明対象のデータに対して微小な変化を与えたサンプルデータとそのモデル予測値を教師データとして、説明したいデータに近いサンプルほど精度良く近似するモデルを作る。解釈可能なモデルで近似することでブラックボックスモデルの挙動を近似的に解釈できる。

学習イメージ図



- 赤い領域と青い領域が元々のモデルの識別範囲。
- 線形モデル(点線)を、サンプル点(赤十字点と青丸点)を参考に学習して、その周辺の近似モデルとする。

深層学習の画像識別モデルに適用した例



(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

- 近似したモデルから予測に強く寄与したピクセルを求めている。
- 左端から、
  (a)元の画像 (b)「エレキギター」の予測に寄与したピクセル
  (c)「アコースティックギター」の予測に寄与したピクセル
  (d)「ラブラドール」の予測に寄与したピクセル

Marco Tulio Ribeiro et al., "Why Should I Trust You?": Explaining the Predictions of Any Classifier, In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016.

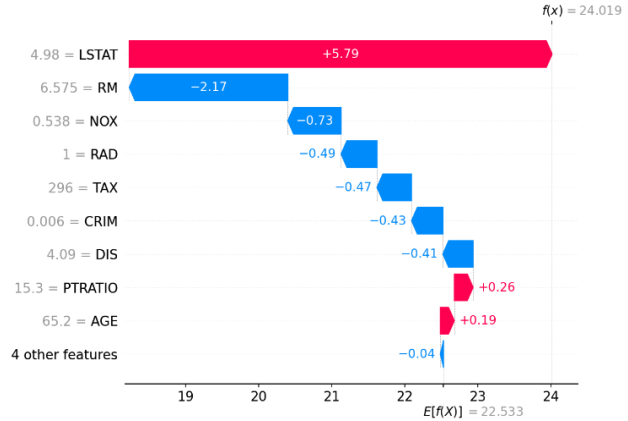
© 2022 The Japan Research Institute, Limited

2.2.7 Surrogate Model - SHAP

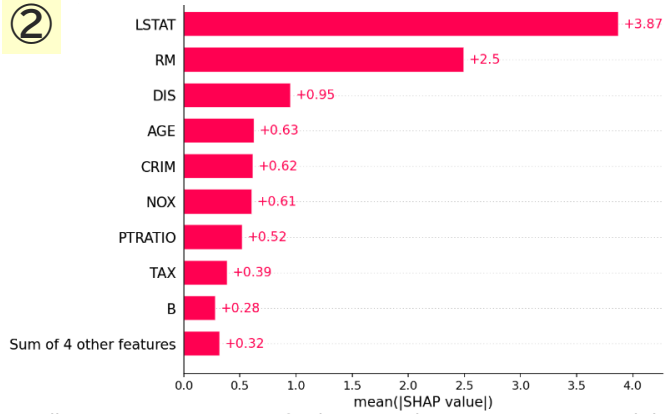
Local Surrogate ー SHAP (SHapley Additive exPlanations)

- SHAPは協力ゲーム理論を応用した手法で、以下のような計算･説明が可能。
  - ① 個々のモデル予測値に対して、各特徴量の寄与度(Shapley値)を計算する。
  - ② 各特徴量のShapley値を平均化することで、モデルのGlobalな特徴量の重要度も計算できる。
  - ③ 特徴量の組み合わせ単位での寄与度も計算可能。

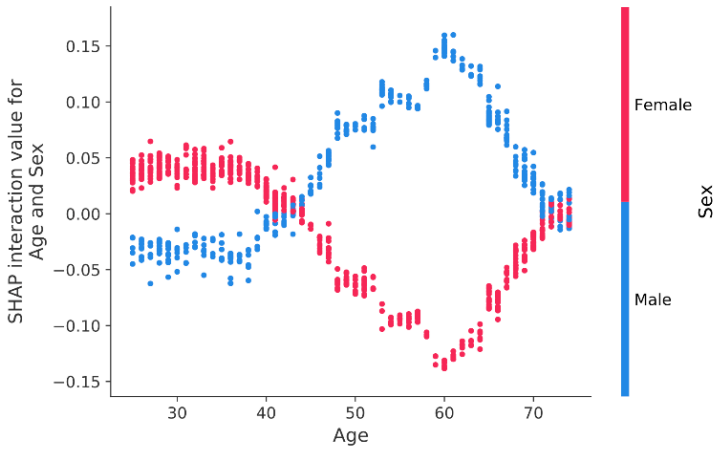
①



②



③



https://github.com/slundberg/shap

(参考)SHAPの基づく理論

- SHAPはゲーム理論におけるShapley値をもとにしている。
- 協力ゲーム理論において、各プレイヤーが連携を行って得られた報酬の総和を各々の貢献度に応じて分配した値をShapley値と呼ぶ。
- この理論を機械学習に拡張すると、特徴量をプレイヤー、モデル予測値を全プレイヤーの報酬和と考えることが出来る(＝モデルの予測値を得るために特徴量がそれぞれどのように寄与したか？)
- SHAPは特徴量のShapley値の線形和が予測値に一致するようにShapley値の推定を行う。

Scott M. Lundberg, Su-In Lee, A Unified Approach to Interpreting Model Predictions, In Proceedings of the 31st International Conference on Neural Information Processing Systems, 2017.

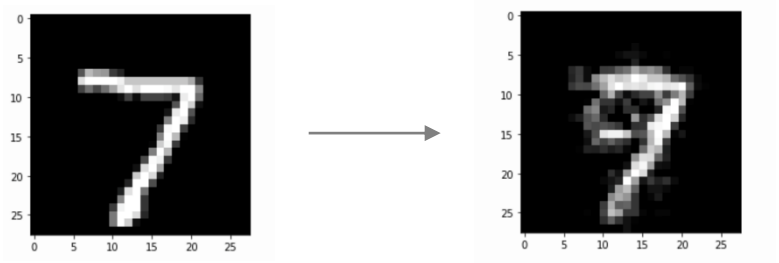
© 2022 The Japan Research Institute, Limited

2.2.8 仮想的なサンプルによる説明可能性

- Counterfactual（反実仮想）な手法では、モデルの予測値に対する解釈として「特徴量の値がもし今と違ったらどうなるか」を仮想的なサンプルを用いて提示する。
- 近しいアプローチで、モデルに誤認識をさせるサンプルを作成する手法としてAdversarial Examplesが知られる。

Counterfactual Explanations

- 対象サンプルデータについて、実際の予測とは異なる望ましい結果になるような似ている点を計算する。
- 対象サンプルとの差分が妥当であるかを評価したり、あるいはユーザーに対するフィードバックとして活用することが出来る。
    - ローン審査に通らなかったが年収が増えれば通る見込み、等



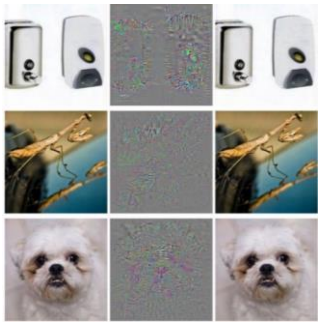
Counterfactuals guided by prototypes on MNIST

数字認識のデータの例。「7」と予測されるサンプルに近い仮想的なサンプルで「9」と予測されるようなを例示する。

Arnaud Van Looveren, Janis Klaise, Interpretable Counterfactual Explanations Guided by Prototypes, In Joint European Conference on Machine Learning and Knowledge Discovery in Databases, 2021.

Adversarial Examples

- Adversarial Examplesはモデルに誤認識させるような、僅かなノイズを加えたサンプルのこと。
    - ニューラルネットワークを騙すようなサンプル生成で話題となった。
    - 結果を変えるという点でCounterfactual Explanationsと似ているが、目的が異なる。
- モデルの脆弱性を明らかにすることで、モデル開発者がモデルの挙動を理解する手助けになる。



左列の画像がオリジナル。中央列のノイズを付加した右列の画像は全て「ダチョウ」として誤認識される。

Christian Szegedy et al., **Intriguing properties of neural networks**, In 2nd International Conference on Learning Representations, 2014.

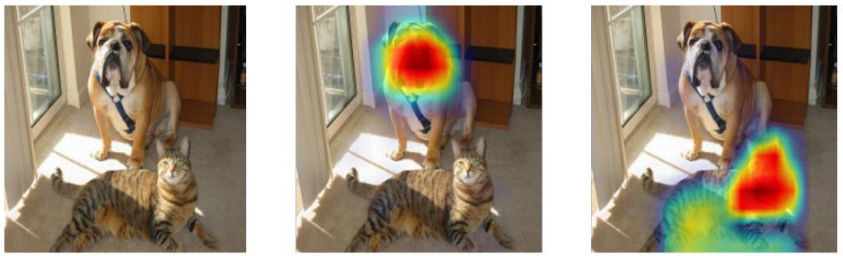
© 2022 The Japan Research Institute, Limited

2.2.9 画像データにおける説明可能性 - Saliency Map

- 画像識別においてどのピクセルが予測に影響を与えているのかを可視化する手法。
- モデルの注目箇所をヒートマップとして表すことができる。

勾配ベースの手法

- シンプルな勾配ベースの手法では、深層学習などの画像識別モデルにおいて、予測値と入力ピクセルの関係を線形なモデルで近似することで線形回帰同様にモデルの挙動を解釈できる*1。
    - 近似のためテイラー展開をすると、各ピクセルの重みがモデルの勾配で表せる。
- 勾配を用いる手法はSaliency Mapを生成する主流のアプローチの一つで、後続の技術開発が盛ん。



Saliency Mapのイメージ*2 。左図が入力画像で、中央図は「ブルマスティフ」の予測に対する注目個所、右図は「タイガーキャット」の予測に対する注目個所が可視化されている。

*1 Simonyan, Karen, Andrea Vedaldi, and Andrew Zisserman. "Deep inside convolutional networks: Visualising image classification models and saliency maps." arXiv preprint arXiv:1312.6034, 2013.

*2 Qinglong Zhang et al., Group-CAM: Group Score-Weighted Visual Explanations for Deep Convolutional Networks, arXiv preprint arXiv:2103.13859, 2021

(参考)Self-Attentionの説明可能性について

- データに基づいてどの特徴量に注目するかを自動的に学習するSelf-Attentionと呼ばれる手法が近年、画像認識の分野においても急速に普及。
    - 元々は自然言語処理の分野で発展した手法。
- 学習された特徴量の注目度合いをそのまま重要度と見なせる。しかし実際には無関係と思われる箇所に注目が発生することが多いなどの問題がある。
- Attentionだけでなく、勾配など他の手法と組み合わせることによって高い説明可能性を実現する提案がされている。

Hila Chefer et al., Transformer Interpretability Beyond Attention Visualization, In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021

© 2022 The Japan Research Institute, Limited

3.市場動向・活用動向

© 2022 The Japan Research Institute, Limited

3.1 取り組み事例の一覧

- 2018年頃から、企業独自のAI倫理／AIガバナンスの原則や指針を打ち出す企業が相次ぐ。
- 近年、原則や指針の打ち出しのみならず、具体的にガバナンスを実現・強化する取り組みが進展している。

カテゴリ	取り組み概要	取り組み企業・組織例
原則/指針の策定 社内組織/委員会の設置	企業独自のAI倫理/AIガバナンスのための原則/指針を策定。 また、AI倫理/AIガバナンスを実現するための体制として、 社内組織や外部専門家を交えた外部委員会を設置。	SMFG、Microsoft、Google、IBM、ソニー、 富士通、NEC、NTTデータ、J.Scoreなど多数
AIの開発・品質評価に係る 標準プロセス等の整備	AI品質ガイドライン、AI開発方法論、AIシステムの倫理評価 などの手順を整備し、案件に適用。	NEC、富士通、NTTデータ
	リスクチェーンモデルの開発。	東京大学未来ビジョンセンター
	AI品質に関するガイドラインを提示。	産業技術総合研究所、QA4AI（AIプロダクト 品質保証コンソーシアム）
AIガバナンスに関する コンサル提供	AI倫理やAI・機械学習モデルに対するコンサルテーション サービス・品質診断サービスを提供。	IBM、日立コンサルティング、ABEJA
公平性に関する個別事例	モデルのバイアス検知や公平性のテストツールを公開。	Google、IBM、PwC、LinkedIn
	公平性に関するチェックリストや対応指針の例示。	Microsoft、AXA
	金融業界における事例。	Zest AI、Scotiabank、HSBC、Amazon、 日本銀行金融研究所、MAS
	データセットのバイアス検知・除去。	Meta、IBMなど多数
説明可能性に関する 個別事例	医療領域、自動運転における適用事例。	理化学研究所＆昭和大学、UC Berkley
	金融業界における事例。	Zest AI、SBI証券、NEC、富士通、LARUS、 オランダ銀行
	鉄道業界の設備保全における劣化進行予測への適用事例。	NEC

青字は次頁以降に詳細解説

© 2022 The Japan Research Institute, Limited

3.2.1 主な商用の製品・ツール｜AI公平性

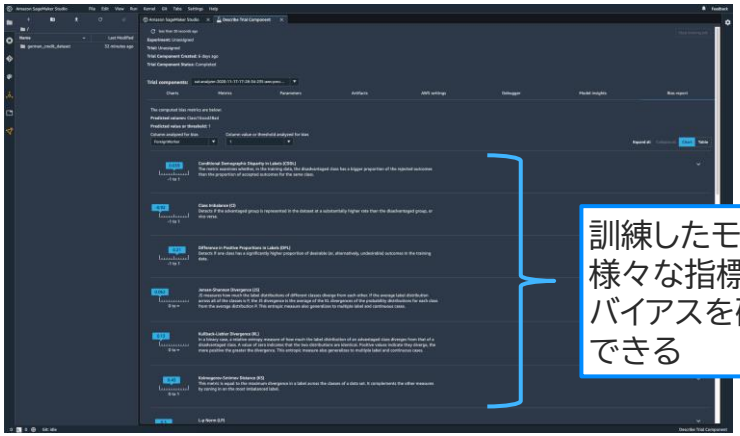
- AI公平性を実現する商用ツールは、クラウドサービスやデータ分析・可視化ソフトウェアに組み込まれて提供されているものが多い。一部、特定のユースケースに特化したツールも存在する。

商用ツールの例

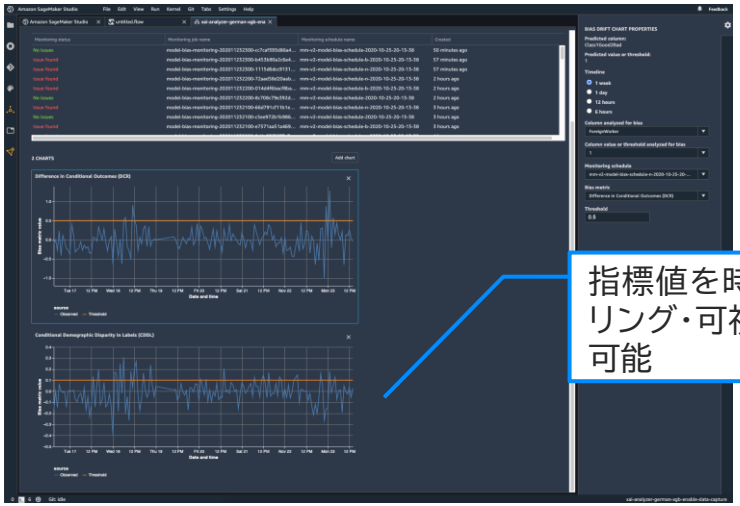
- AI公平性に関する商用ツールの例として以下表の通り。これらのツールには、基本的に、AI公平性に関連する定義・指標、バイアスを軽減するための学習アルゴリズムが実装されている。
- センシティブ属性を指定すると、これらの属性にバイアスが存在するかどうかを検出できる。視覚的なレポートを提供しており、バイアスを修正する手助けをしてくれる。
- また、実稼働システムのバイアスを監視し、予め決めた閾値を超えるとアラートを発出するなども可能なツールがある。
- 一方、「Spellcheck for Bias」は特定のユースケースに特化したツールである。映画やテレビの台本、原稿、広告概要などのテキストデータを分析し、性別・人種・LGBTQIA+・障害・年齢・体型などの6つの人物に関する表現を抽出することができる。加えてステレオタイプ、暴力、差別などの属性の分析も可能である。

名称	販売元企業
Azure Machine Learning	Microsoft
Amazon SageMaker Clarify	AWS
DataRobot	DataRobot
Fiddler	Fiddler AI
Spellcheck for Bias	Geena Davis Institute

（参考）Amazon SageMaker Clarifyのイメージ



訓練したモデルに関して、様々な指標値を計算し、バイアスを確認することができる



指標値を時系列でモニタリング・可視化することも可能

https://aws.amazon.com/jp/sagemaker/clarify/

© 2022 The Japan Research Institute, Limited

3.2.2 主な商用の製品・ツール｜XAI

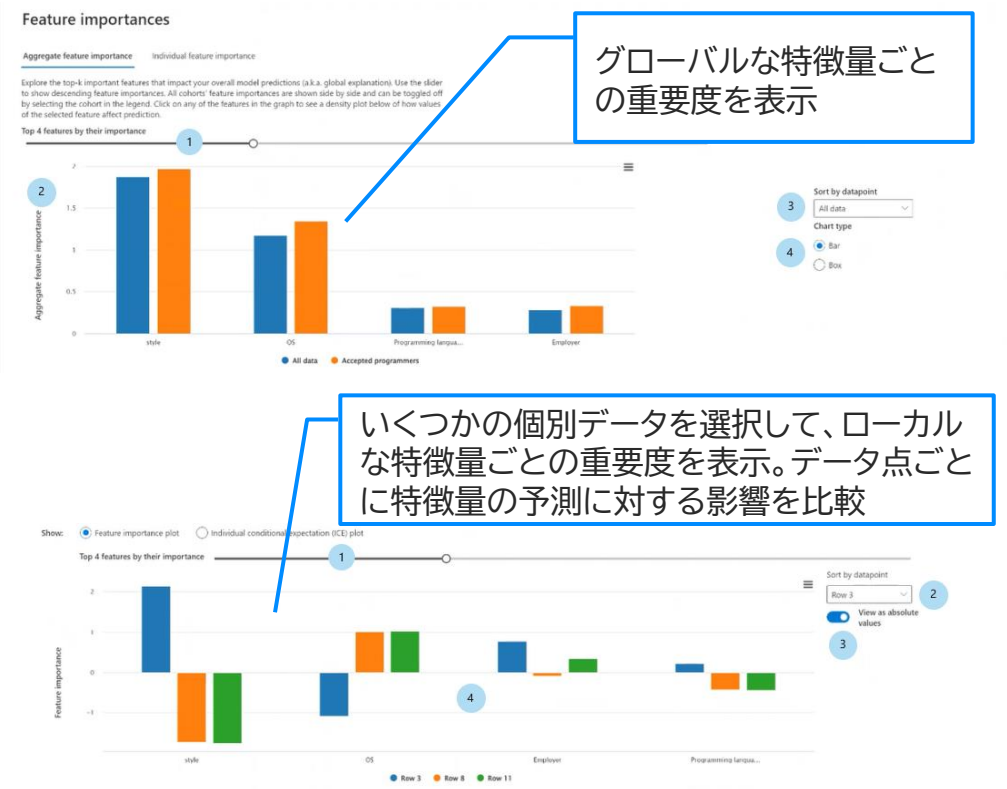
- 説明可能AIの機能がクラウドサービスに組み込まれて提供されることが多い。どのプラットフォームでも表形式のデータの説明可能性には対応しているが、画像などへの対応状況などに差がある。
- AWSやAzureのようなクラウド上のプラットフォームでは機能追加のサイクルも早く、今後も機能の拡充が進むと予想。

商用ツールの例 （※）2022年9月調査時点の情報で記載

- XAIに関する商用ツール例は下表の通り。
- 基本的に、表形式データに対するモデル説明可能性を提供しており、特徴量の重要度を可視化する機能が提供されている。使われている説明手法は様々だが、LIMEやSHAPをベースとした手法が多い。
- 表形式データ以外では画像データに対する説明手法が実装されているプラットフォームもある。例えばAWSでは画像分類、物体検知に対する説明を、GCPでは画像分類に対する説明を提供している。
- IBM Cloudの場合は、AWS/Azure/GCPと異なり、パブリック・クラウドとプライベートクラウドの両方でハイブリッドに対応可能な点が特徴となっている。

プラットフォーム	機能名称
AWS	SageMaker Clarify
Azure	Responsible AI dashboard
GCP	Vertex Explainable AI
IBM Cloud	Watson OpenScale

（参考）Responsible AI dashboardのイメージ



https://learn.microsoft.com/en-us/azure/machine-learning/how-to-responsible-ai-dashboard

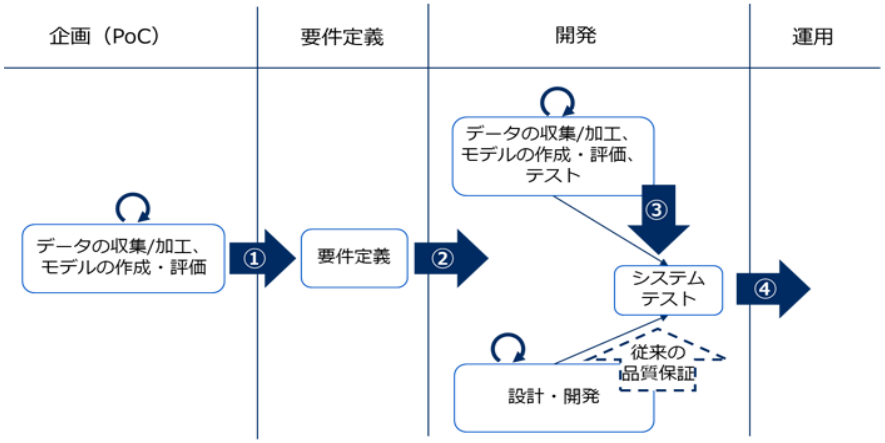
© 2022 The Japan Research Institute, Limited

3.3.1 代表的な取り組み事例｜標準プロセス等の整備

- AIモデルの公平性や透明性を確保するための取り組みとして、AIシステム開発・運用時の標準プロセスや、リスク軽減のためのフレームワークの策定が進む。
- 日本国内ではNEC社が2019年より「AI品質ガイドライン」を策定し、運用しているほか、東京大学は、AIサービスが引き起こすリスクと低減策を検討する「リスクチェーンモデル」を公表している。

AI品質ガイドライン（NEC）

- NECは、AIシステム特有の品質担保を目的に「NEC AI品質ガイドライン」を策定し、運用している。
- AIシステム開発において重要な①システムの企画（PoC）②データの収集/加工、③モデルの作成/評価/テスト、④システム運用の4フェーズごとにチェック項目を設定して具体的基準を策定している。
- 各フェーズ間を移行する際にガイドラインに従ってチェックすることで、リスクを早い段階で防止する。

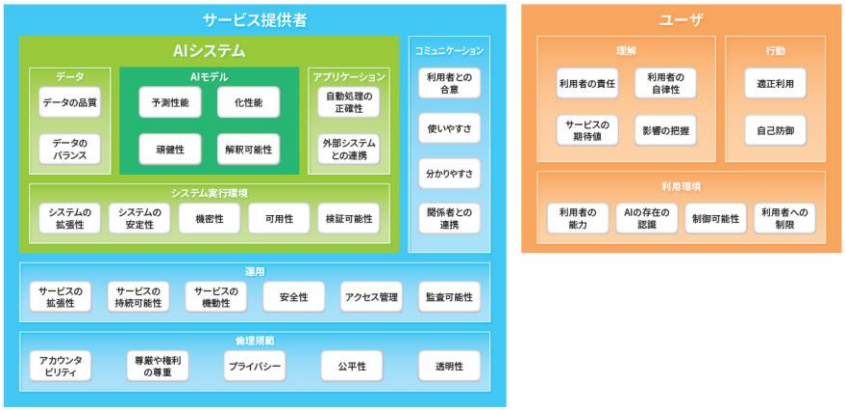


NEC、「AI品質ガイドライン」を策定し、AIシステムの構築・開発に適用, 2019年12月, https://jpn.nec.com/press/201912/20191210_02.html

リスクチェーンモデル（東京大学）

- 東京大学未来ビジョンセンターは、様々な原則やガイドラインが公開されているなか、それらの原則を実践に移す一手法として、リスクチェーンモデルが公開されている。
- AIサービスが引き起こす重要なリスクシナリオを検討し、リスク低減策およびステークホルダー毎の役割を整理できる。
- リスクチェーンモデルの公開サイト*1では、ケース事例が公開されており、検討の参考となる情報も整っている。

*1 https://ifi.u-tokyo.ac.jp/projects/ai-service-and-risk-coordination/



東京大学未来ビジョンセンター、リスクチェーンモデル（RCModel）ガイド Ver1.0, 2021年6月, https://ifi.u-tokyo.ac.jp/wp/wp-content/uploads/2021/07/RCM_210705.pdf

© 2022 The Japan Research Institute, Limited

3.3.2 代表的な取り組み事例｜AI公平性

- 一部の企業・団体では、公平性を実現するためのチェックリストやツールを公開している。
  - **Microsoft** ： 公平性を実現するためのチェックリストを公開。
  - **LinkedIn** ： 大規模なデータを扱うことができ、スケーラビリティのある公平性のためのツールを開発・公開。

AI Fairness Checklist(Microsoft)

- AIシステムの開発に携わる48人の実務家と共同設計して作成。
- 技術的側面のみではなく、社会技術(sociotechnical)的な要素を取り入れたチェックリストであることが特徴。
- チェックリストをチーム・組織のワークフローと整合させ、組織文化に支えられた状態にすることが重要であると指摘。

AI Fairness Checklist

The items in this checklist are intended to be used as a starting point for teams to customize. Not all items will be applicable to all AI systems, and teams will likely need to add, revise, or remove, items to better fit their specific circumstances. Undertaking the items in this checklist will not guarantee fairness. The items are intended to prompt discussion and reflection. Most items can be undertaken in multiple different ways and to varying degrees.

Envision

Consider doing the following items in moments like:

- Envisioning meetings
- Pre-mortem screenings
- Product greenlighting meetings

1.1 Envision system and scrutinize system vision

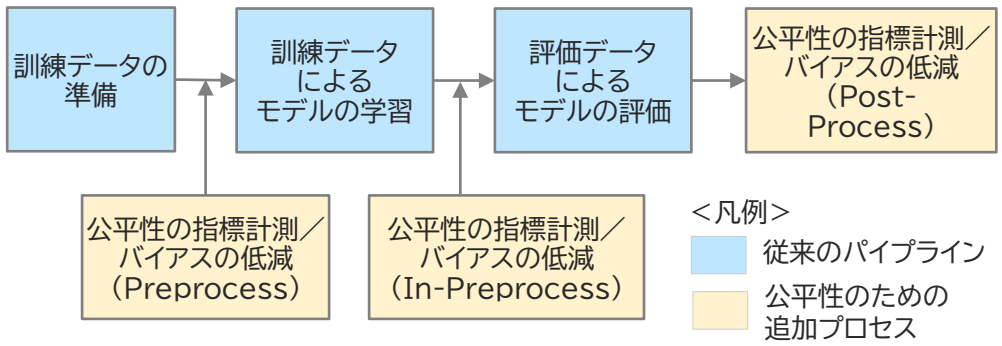
1.1.a Envision system and its role in society, considering:

- System purpose, including key objectives and intended uses or applications
  - Consider whether the system should exist and, if so, whether the system should use AI
- Sensitive, premature, dual, or adversarial uses or applications
  - Consider whether the system will impact human rights

Microsoft, AI Fairness Checklist,
https://www.microsoft.com/en-us/research/project/ai-fairness-checklist/ より抜粋

LinkedIn Fairness Toolkit(LinkedIn)

- LinkedIn社は、大規模な機械学習システムに組み込むことが可能な、LiFT(LinkedIn Fairness Toolkit)を開発し、公開。
- モデル学習前・学習中・学習後それぞれのバイアス検知・軽減の機能を提供しており、既存の機械学習パイプラインから柔軟に呼び出すこと可能。
- 複数ノードからなる計算環境でも稼働させることが可能で、大規模なデータセットでも処理可能であることが特徴。



「Vasudevan Sriram et al, LiFT: A Scalable Framework for Measuring Fairness in ML Applications, CIKM '20, 2020」のFigure1をもとに作成

© 2022 The Japan Research Institute, Limited

3.3.3 代表的な取り組み事例｜AI公平性

- モデルの公平性を確保する取り組みとして、公平性に配慮した学習用データセットの構築が進む。
- 特に、顔認識モデルでは、人種によって顔認識の精度が異なることが指摘されており、人種差別による警察の誤認逮捕などの問題を引き起こす可能性が指摘されている。
- そのため、Meta社は年齢・性別・見た目の肌の色・照度などの公平性に配慮したデータセットを公開している。

Gender Shades

- 人種・性別によって顔認識の精度が異なることを指摘したBuolamwini氏らの有名な研究が存在する。
- 3つの商用の顔認識ソフトウェア(APIなど)を対象に、画像中の人物の性別判別を実施(下表は肌の色の印象・性別ごとの判別エラー率をまとめたもの)。
- 肌の色が暗い人物の場合にエラー率が高くなり、特に女性の場合にエラー率が高くなることが報告されている。

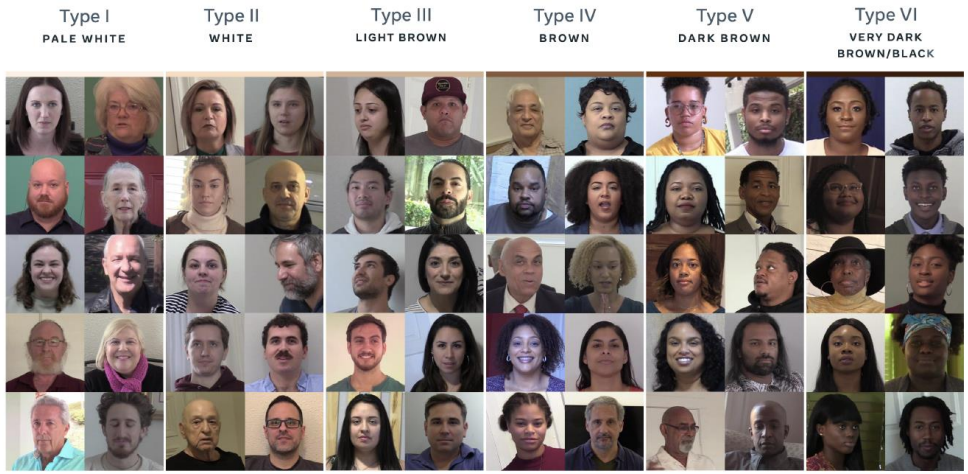
モデル	肌の色が暗い		肌の色が明るい	
	女性	男性	女性	男性
Microsoft	20.8	6.0	1.7	0.0
Face++*1	34.5	0.7	6.0	0.8
IBM	34.7	12.0	7.1	0.3

*1 中国に本社を持つコンピュータビジョン技術に強みを持つ企業

Joy Buolamwini et al, Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification, Proceeding of the 1st Conference on Fairness, Accountability and Transparency, 2018 のTable4をもとに作成

公平性に配慮したデータセットの構築(Meta)

- Metaは、いくつかの顔の属性に渡ってAIモデルの頑健性を評価するためのデータセットを構築。画像処理(特に顔属性の分類)や音声理解などへの応用が期待される。
- データセットは3,011人の被験者が参加し、45,000以上のビデオから構成され、年齢・性別・見た目の肌の色・照度についてアノテーションされている。



Caner Hazirbas et at, Towards Measuring Fairness in AI: the Casual Conversations Dataset, 2021, arXiv preprint, arxiv:2104.02821, https://arxiv.org/abs/2104.02821

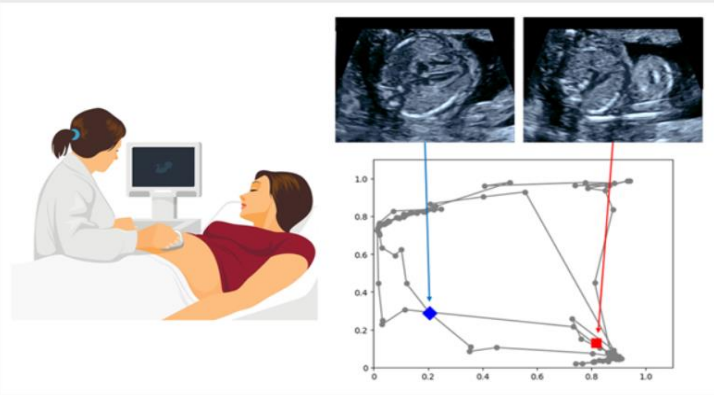
© 2022 The Japan Research Institute, Limited

3.3.4 代表的な取り組み事例｜XAI

- XAIの技術発展に伴い、様々な業種において活用の検討が進んでいる。特に、ミッションクリティカルな領域では、ブラックボックスAIの判断根拠を人間が確認するため、説明可能性に注目した研究が進む。
- 医療分野では検査者の診断を支援する事例、自動運転に関する研究では操作の自動判断について判断根拠を提示する事例がある。

超音波スクリーニング診断（理研、昭和大学）

- 胎児心臓超音波スクリーニング診断*1において、理研および昭和大学の共同研究グループがAIの判断根拠を可視化して検査者を支援する技術を開発。
- 異常所見の有無の判定時に、根拠となる診断部位の検出結果を従来よりも明確に提示でき、実際に検査者が根拠を参考にすることで、スクリーニング精度が向上することを確認。

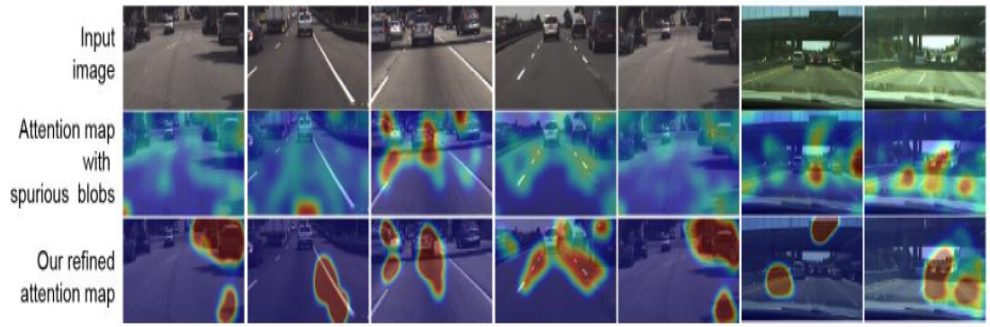


グラフチャート図を用いた超音波画像診断支援

理化学研究所 2022年3月22日 プレスリリース
https://www.riken.jp/press/2022/20220322_2/index.html

自動運転におけるステアリング操作（DARPA）

- 自動運転の操作における、注目個所の可視化と操作理由の言語化の研究。
- UC Berkleyの研究グループが実施したDARPAのXAIプロジェクトの一部。
- 自動運転におけるステアリング操作について、操作の根拠となった入力画像上の注目箇所のより良い可視化と言語化を試みている。



Kimら, “Show, Attend, Control, and Justify: Interpretable Learning for Self-Driving Cars”, 2017

*1 先天性心疾患の早期発見のため、胎児心臓を観察する超音波検査

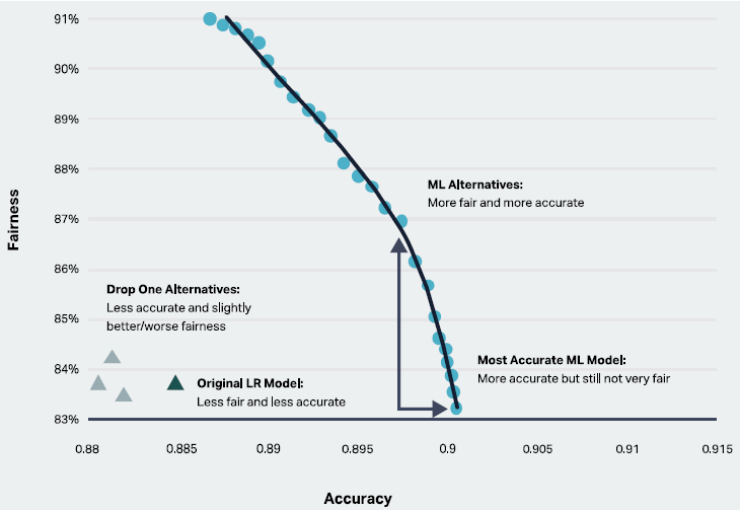
© 2022 The Japan Research Institute, Limited

3.4.1 金融分野における取り組み事例｜AI公平性・XAI

- 金融分野で公平性・説明可能性の観点で先進的な取り組みを進めている企業として Zest AI がある。
- Zest AI は、AIを活用したローン審査モデル（ローン受付時の信用スコアの算出）を開発し、金融機関に提供している。公平なローン審査を実現することにより、従来はローンを借りづらかった顧客層に対してもローン対象者を増やすことができたと報告されている。

公平性

- 規制当局等への報告などを意識して、ソフトウェアの中で、Disparate Treatment（使用者による個別の差別的な取扱い）やDisparate Impact（職場の制度や慣行の差別的影響）を自動的に分析する機能が提供されている。
- また、モデルの性能と公平性の基準値のバランスを選択できるようにするため、適切なモデルをサーチする機能も提供されている（下図）。

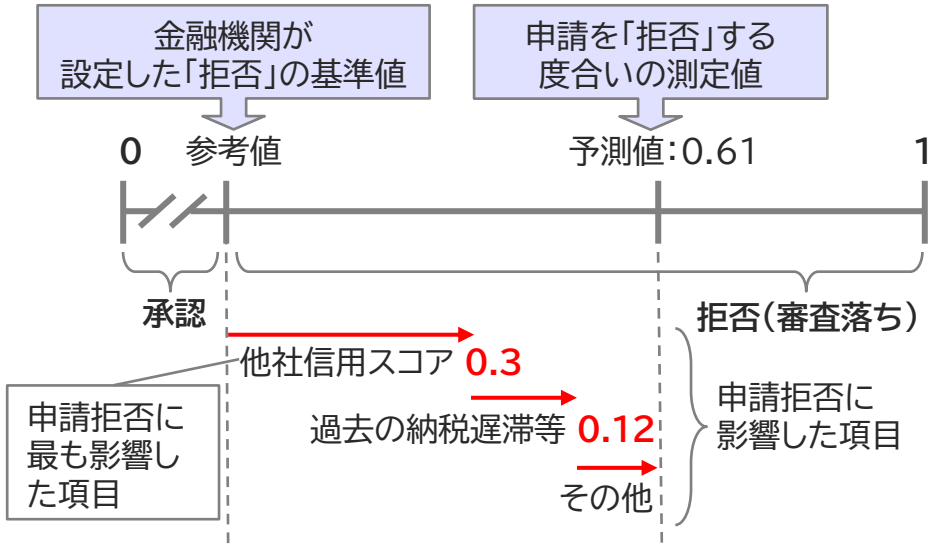


ZestAI, Five Building Blocks for Compliant AI-driven Lending
https://www.zest.ai/resources/five-building-blocks-for-compliant-ai-driven-lending

説明可能性

- スコアリングに寄与する主要な要素を特定し表示が可能であり、なぜローン審査が落ちたと判断されたのかを説明することが可能。
- 説明にはゲーム理論を活用したShaplay値をベースにカスタマイズした値が使用されているとのこと。

ローン審査を拒否する場合のイメージ



© 2022 The Japan Research Institute, Limited

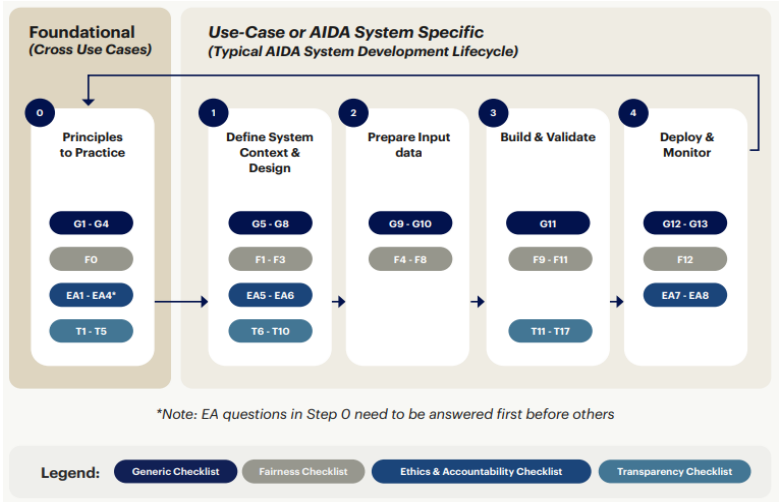
3.4.2 金融分野における取り組み事例｜AI公平性・XAI

- シンガポール金融管理局(MAS*1)が主導するVeritasコンソーシアムは、2022年2月に、金融機関によるFEAT*2の原則に関する、評価方法を詳述したホワイトペーパーをリリース*3。
- AIDA*4システムの開発ライフサイクルに採用するべきFEATに関するチェックリストと、①公平性、②倫理・説明責任、③透明性の3つの観点での評価方法論が整理されている。

*1 MAS : Monetary Authority of Singaporeの略　*2 FEAT : Fairness, Ethics, Accountability, Transparencyの略
*3 https://www.mas.gov.sg/news/media-releases/2022/mas-led-industry-consortium-publishes-assessment-methodologies-for-responsible-use-of-ai-by-financial-institutions
*4 AIDA : Artificial Intelligence and Data Analysisの略

FEATチェックリスト

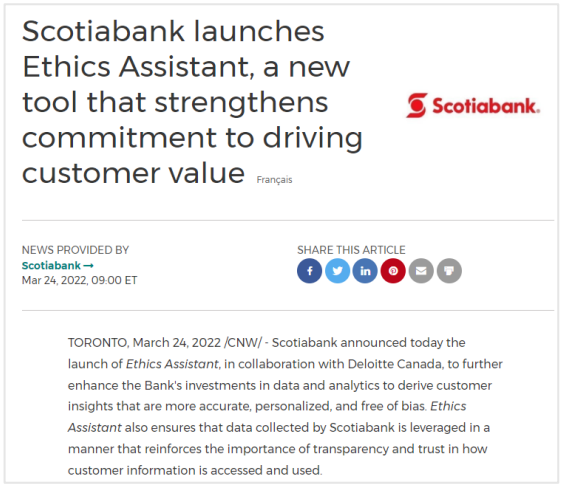
- AIシステムの利用用途に依らず共通して適用するチェック項目(Fundamental)と、個別システムの利用用途ごとに適用可否を検討するチェック項目(ライフライクル別)に分かれている。
- また、チェック項目は公平性など観点別にも分かれている。



Veritas Document 3 : FEAT Priciples Assessment Methodology, https://www.mas.gov.sg/-/media/MAS-Media-Library/news/media-releases/2022/Veritas-Document-3---FEAT-Principles-Assessment-Methodology.pdf

Ethics Assistant(Scotiabank)

- AIモデル開発者が、開発プロセスの早い段階からモデルの倫理的な配慮について考えられるように「Ethics Assistant」と呼ばれるツールを開発*5。 *5 デロイト社のTrustworthy AI Impact Assessment Toolを利用
- 説明責任、公平さなど、考えうる網羅されたリスクに対して、懸念事項に対処するための実践的なガイダンスを得ることができる。


Scotiabank launches Ethics Assistant, a new tool that strengthens commitment to driving customer value

Français

Scotiabank

NEWS PROVIDED BY
Scotiabank
Mar 24, 2022, 09:00 ET

SHARE THIS ARTICLE

TORONTO, March 24, 2022 /CNW/ - Scotiabank announced today the launch of *Ethics Assistant*, in collaboration with Deloitte Canada, to further enhance the Bank's investments in data and analytics to derive customer insights that are more accurate, personalized, and free of bias. *Ethics Assistant* also ensures that data collected by Scotiabank is leveraged in a manner that reinforces the importance of transparency and trust in how customer information is accessed and used.

https://www.newswire.ca/news-releases/scotiabank-launches-ethics-assistant-a-new-tool-that-strengthens-commitment-to-driving-customer-value-849190964.html

© 2022 The Japan Research Institute, Limited

4.展望・考察

© 2022 The Japan Research Institute, Limited

4.1 AI公平性・XAIに関する技術的な課題、展望・考察

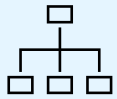
- 公平性・説明可能性を実現するツールも登場し、一部技術は実用化の段階に移っている。しかし、現時点でAIを社会実装する上で求められるニーズを全て満たしているわけではなく、また、デファクトと言えるような手法も存在しない。
- 今後も基礎研究～実用化の全フェーズで研究開発が進展していくと考えられる。

	主な課題	展望・考察
AI公平性	公平性の定義について問題点が指摘されている。 • 類似の公平性定義が多数あり、システム企画/開発者にとって、採用すべき指標が分かりにくい。指標間の関係性も分かりにくい。 • 公平性は社会技術的(socio-technical)な観点が必要とされ、本来は複雑な問題である。一方で、公平性の定義は数式としてシンプルに定式化され過ぎている。 • 現状、社会的な合意がなされているとは言い難い。	• 公平性を測る統一的な指標の探索が続く。指標の使い分けに関するガイドライン等も整備され、システム企画/開発者がより使いやすい環境が整う。 • 因果ベースの定義や、個々の参加者ごとに異なる指標に関して不利益を防ぐゲーム理論ベースの定義など、より発展的なアプローチが模索されていく。
	• 研究対象のデータ・問題設定に偏りがある。既存研究の多くがテーブルデータを対象とし、分類問題となっている。	• 自然言語(特に言語モデル)に対する公平性の研究が進展。 • 回帰問題や、より難しい問題設定での研究も進展していくと考えられる。
XAI	• 手法によりAIモデルの性能とのトレードオフがある。解釈可能モデルなどモデル自体が解釈しやすいものは、複雑なモデルと比べ性能面で劣りやすい。	• 性能を犠牲にしない方向の技術開発が引き続き進む。適用する技術について適宜見直す必要がある。
	• 適用するタスクや対象とするユーザによって、望ましい説明可能性が異なる。 • 説明が期待したものでない場合、信頼性が損なわれることがある。	• 具体的なユースケースやユーザを想定した詳細な評価が必要。 • 手法の性質を理解して使い分けるユーザ側の理解も必要。 • AIとよりよく協調するため、人間の認知も含めた研究が進む。
	• 異なるAIモデルで同程度の性能を達成できることがある。これにより同じ予測に対して複数の異なる説明が得られるケースがあり、妥当でない説明を使ってしまう可能性がある。	• 妥当でない説明性の検知可能性/検出技術への期待。 • 妥当でない説明を意図的に用いて何らかの利益を得ようとするプレイヤにペナルティを課す制度設計。

© 2022 The Japan Research Institute, Limited

4.2.1 AI公平性・XAIの実現に向けた推奨事項

- 企業が、AIの公平性・説明可能性の取り組みを進めていく際の推奨事項は、以下の通り。

	推奨事項	実施者	概要
共通	新技術、ツールの調査・技術評価・研究開発 	ユーザ部門 ↑ 技術評価 システム部門 調査・研究開発	• AIに関する法規制はまだ実運用に至っておらず、各国の標準・ガイドライン等も乱立する可能性が排除できないため、当面は過度に対応する必要性は低い。 • しかしながら、AIを社会で安全安心に利用してもらうという本来の目的を実現するため、先進企業に倣って、現時点から、技術面における公平性、説明可能性の準備、事例調査等の取り組みを開始する必要がある。
	AI倫理・ガイドラインの実行体制の整備 	ユーザ部門 ⇔ 協働体制 システム部門	• 近年、AI倫理を重要だと考える経営層の割合は急増。一方、企業はAI倫理への取り組み意欲を高めているが、実践が追い付いていない。 • まずは、AIの技術的な理解に加えて、AIの管理運用を組織的にかつ効率的に行うための組織・ガバナンス体制の構築が必要である。
	システム企画段階から検討を開始（P.41-42） 	ユーザ部門 ↑ 技術支援 システム部門	• AIシステムの公平性・説明可能性をどこまで実現するのかは、システムが提供するサービス内容や、システムの出力が顧客等に与える影響やリスクに基づいて判断する必要がある。 • そのため、AIシステムの開発完了時に公平性・説明可能性を検証するのではなく、システムの企画段階から検討を開始することが重要である。
公平性	公平性に関する社会的な合意形成（P.43） 	ユーザ企業 公的機関 一般市民	• 公平性については技術的な側面だけでなく、社会技術的な観点が必要となる。 • そのため、自社にとってリスクが高いとされるユースケースを優先的に、業界横断で、AIシステムの公平性の定義や指標に関する議論・検証を重ね、社会的な合意形成を図ることが必要。 • また、合意形成を図るための、オープンな仕組み・場づくりも必要となる。
説明可能性	適切な手法、アプローチの見極め（P.44） 	ユーザ部門 ⇔ 協働検討 システム部門	• 説明可能性はステークホルダーの理解を促すものであり、AIシステムの提供する機能やステークホルダーの立場・役割によって適切な手法が異なる（万能なXAI手法はない）。 • そのため、担当者はXAIに関する各手法のアプローチや、説明する対象を理解して使い分ける必要がある。

© 2022 The Japan Research Institute, Limited

4.2.2 システム特性に応じたAI公平性・XAIの実現

- AIシステムの公平性・説明可能性をどこまで実現するのかは、システムが提供するサービス内容、システムの出力が顧客等に与える影響やリスクに基づいて判断する必要がある。
- そのため、AIシステムの開発完了時に公平性・説明可能性を後付けで検証するのではなく、システムの企画段階から検討を開始することが必要である。
- 特に、公平性については、社会での捉え方が時間とともに変化するため、AIモデルの開発フローを柔軟に変更したり、公平性の概念を柔軟に組み込めるように予め準備しておくことが望ましい。

AIシステム利用に関する規制

- システムによって求められる公平性や透明性、説明責任は異なる。欧州AI規制ではリスク分類に基づき、提供者の義務が異なる。
- AIシステムの運用について、規制やガイドラインを参考に対応方針を検討する必要がある。

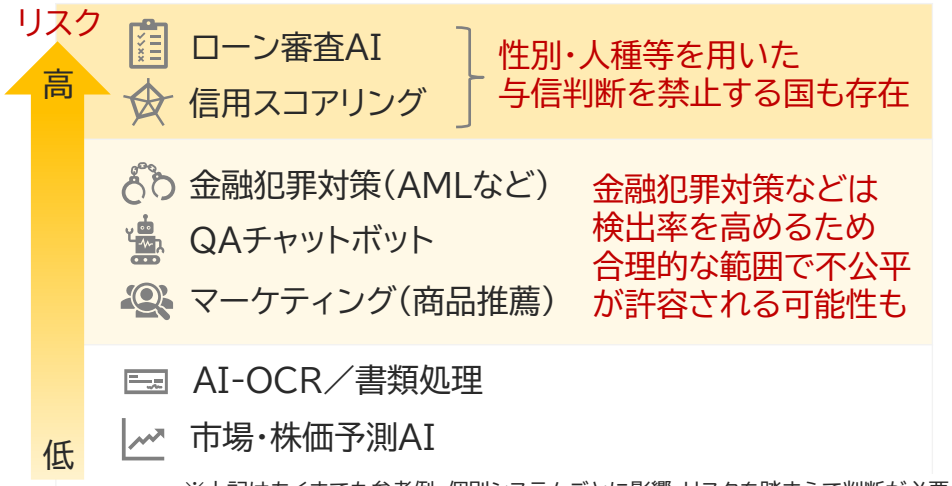
欧州AI規制法案におけるリスクの考え方と分類の概要

レベル	対象	適用内容
受容できないリスク	基本的人権に反するもの • 人々の意識を超えたサブリミナルな手法を用いて人を操作するもの • 公的機関のソーシャルスコアリング（差別的な結果及び一定のグループの排除を招くおそれ）	原則禁止
高リスク	• 雇用、教育、医療、法執行などの分野で個人の重大な利益に係る意思決定 • 重要なインフラ、生体認証などで用いられる安全性や基本的人権に悪影響を与えるAIシステム	• リスク管理プロセスの構築 • 透明性及びユーザへの情報提供 • 人間による監視
限定的なリスク	自然人と相互作用するシステム • チャットボット	人とAIシステムの相互作用があることを通知する

（参考）金融業界におけるAIシステム利用のリスク

- 下図は金融業界で活用されるAIを例に、公平性や透明性の観点でリスクを整理。
- 個人の利益に影響する「ローン審査」などがハイリスクであり、「株価予測」など、公平性と関連の薄いAIはローリスクと想定される。

金融機関で活用されるAIの例



※上記はあくまでも参考例。個別システムごとに影響・リスクを踏まえて判断が必要

© 2022 The Japan Research Institute, Limited

4.2.3 AI実装の各段階におけるAI公平性・XAIの評価項目例

- 経済産業省「AI原則実践のためのガバナンス・ガイドライン」において、公平性・説明可能性の観点における、AIガバナンス・ゴールとの乖離評価例が挙げられている（下表はその一部）。
- 評価項目はあくまでも例であり、各AIシステムの事情に応じて修正・取捨選択が必要である。

段階・分類		評価項目例	
		AI公平性	XAI
企画・設計段階	置かれている状況	AIシステム開発者及び運用者は、AIシステムに求められる公平性を把握しているか。 ・開発・運用しようとしているAIシステムや類似のAIシステムの公平性に関するインシデント事例を調査したか。	AIシステム開発者及び運用者は、AIシステムに求められる身体、精神、財産等への悪影響を把握しているか。 ・開発者は、AIシステムが身体、財産等に影響を与えうる場合、標準的な実務などに照らして許容されるリスクを評価したか。
	設計全般	AIシステム開発者は、AIシステムの公平性に関する課題に対処したか。 ・他国へのAIシステム提供を想定する場合、その国の規制、慣習、慣行等を理解する人を開発チームやレビュー役に加えたか。	AIシステム開発者は、AIシステムの機能、効果について、AI以外のシステムと比較しながら、AIシステム運用者とすりあわせをしたか。 ・開発者は、精度と説明可能性等のトレードオフを考慮して、運用者がAIシステムに期待する機能、効果を提供できない場合には、その旨を説明し、必要に応じて代替策を提案したか。
開発段階	データ	AIシステム開発者は、データセットの設計にあたり、特定の社会属性に基づく不当な差別を維持・助長しないよう配慮したか。 ・特定の社会属性に基づく不当な差別を維持・助長するようなデータセットを用いないようにしたか。	-
	モデル・システム	AIシステム開発者は、開発しようとしているAIシステムの公平性を確保したか。 ・AIシステム開発者は、公平性の定義や指標を調査し、適当な場合には、公平性の定義や指標を定め、公平性を客観的に評価したか。	AIシステム開発者は、開発しようとしているAIシステムの説明可能性に配慮したか。 ・AIシステム開発者は、全ての出力について、人間が理解できるような一定の説明を加えることができるか否か確認したか。
運用・モニタリング		AIシステム運用者は、AIシステム開発者によるAIシステムの公平性に関する課題への対処の内容を理解したか。 ・AIシステム運用者は、許容される範囲を超えた出力データに対する警告を発する措置を講じた場合に、警告の意味を理解したか。理解できない場合に、開発者に問い合わせ、疑問を解消したか。	AIシステム運用者は、AIシステム利用者に対する説明責任を果たしているか。 ・AIシステム運用者は、AIシステム開発者の支援を仰ぎながら、全ての出力について、要請があった場合に、人間が理解できるような一定の説明を加えることができることを確認したか。

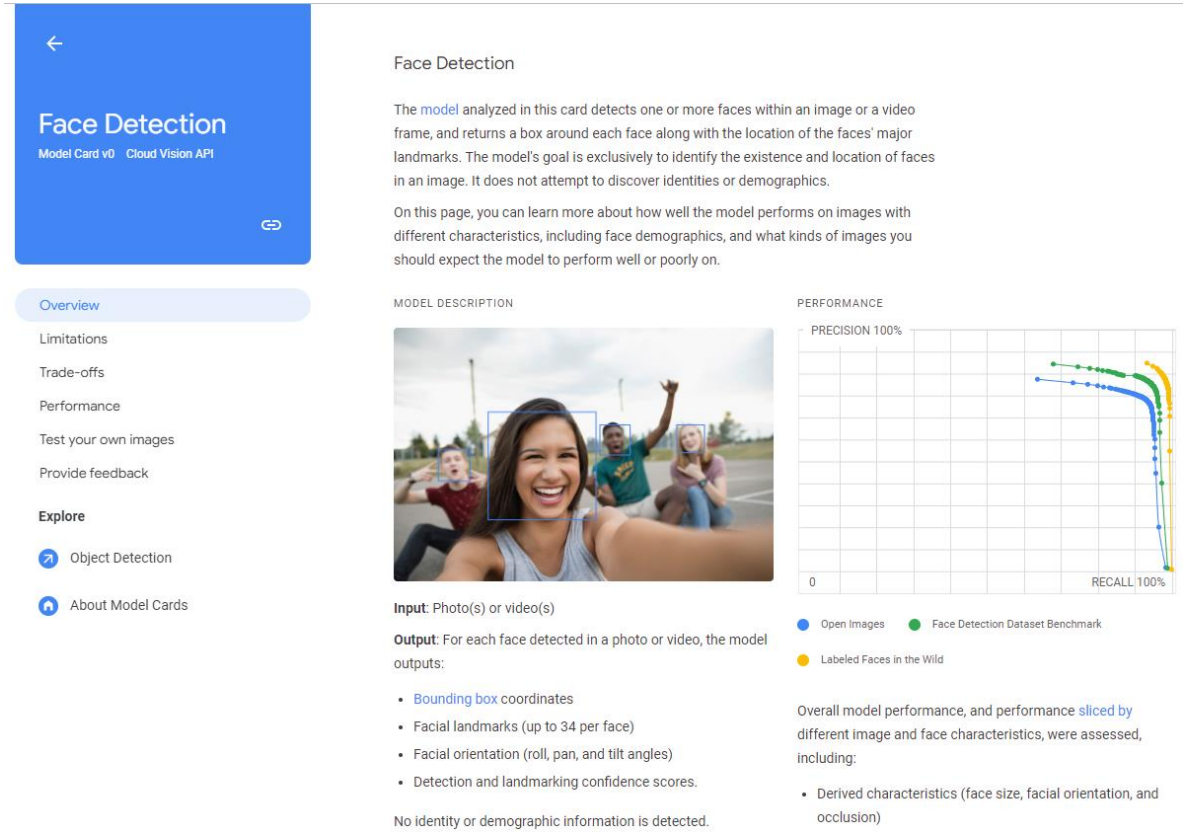
経済産業省「AI原則実践のためのガバナンス・ガイドラインver.1.1」のうち「別添2（AIガバナンス・ゴールとの乖離を評価するための実務的な対応例）」をもとに作成

© 2022 The Japan Research Institute, Limited

4.2.4 公平性に関する合意形成を得るためのツール｜モデルカード

- 公平性に関して社会的な合意を形成していくため、提供するAIシステムが持つ特性をオープンに説明できるように準備することが望ましい。そのための参考となるフレームワークとしてモデルカードがある。
- モデルカードでは、モデルのタスク、想定利用方法、評価指標、訓練・評価データ、倫理上の考慮事項など様々な観点で、モデルの性質をカード形式で整理する。

モデルカードのサンプル



Google Cloud Model Cards（https://modelcards.withgoogle.com/face-detection） 2022/9/20アクセス

（参考）倫理上の考慮事項

モデルカードの論文*1では、倫理上の考慮事項として以下の観点が挙げられている。

観点	内容
Data	センシティブなデータを使っているか？
Human Life	人間の生命・生活の根幹に関わる事柄を決定するためのモデルか？
Mitigations	モデルの開発プロセス中にどのようなリスク軽減策を講じたか？
Risks and harms	モデルの利用によってどのようなリスクがあるか？起こり得る損害とその大きさの特定を試みる。不明な場合はそのことも記載する。
Use Cases	特に問題となる既知のモデルのユースケースは存在するか？

それ以外にも、開発者以外の組織がモデルをレビューした実績など、他の対応事項もあれば記載すると良い。

*1 Margaret Mitchell et al, Model Cards for Model Reporting, arXiv:1810.03993, https://arxiv.org/pdf/1810.03993.pdf

© 2022 The Japan Research Institute, Limited

4.2.5 ユーザごとの目的と用いるXAI手法(金融機関の例)

- 説明可能性は、AIシステムの提供する機能やステークホルダーの立場・役割によって、適切な手法が異なる(下記の表は、金融機関の例)。
- 担当者はXAIに関する各手法のアプローチや説明する対象を理解し、手法がステークホルダーの要望や規制・ガイドライン等の要求に適しているかを検討して使う必要がある。

金融ドメインにおけるロールと目的の例

ロール	目的	XAI利用例	
モデル開発者 データサイエンティスト	説明可能性と精度を評価することによって、ユースケースに応じたモデル開発を行う。	Local Surrogate, 解釈可能モデル	• 特徴量が予測に与える影響を説明できるかを確認したり、得られた知見から特徴量生成して、モデル開発を行う。 • 説明可能性を優先するユースケースの場合、解釈可能モデルを開発する。
リスク担当者・ 内部監査・銀行幹部	モデルのロバスト性に関するテストを実施することで、モデルに内在するリスクを明らかにし、対策を講じる。	Adversarial Examples, Global Surrogate	• Adversarial Examplesを用いてモデルの脆弱性を明らかにし、対策を講じる。 • Global Surrogateでブラックボックスモデルの挙動の妥当性を検証する。
銀行従業員	課題解決のための知見をモデルから得ることで、業務を改善する。	PDP, Local Surrogate	• 重要な特徴量の可視化や個々の予測に対する説明から得られた知見に基づいて業務改善のための施策を立案する。
規制当局・監査法人	公正な貸し付けや顧客保護などのコンプライアンス遵守の認定を行う。	Global Surrogate	• 重要度の高い特徴量を抽出し、これらを用いることが倫理違反や人権侵害に該当しないことを確認する。
銀行顧客	モデルによって行われる意思決定に対し、改善できる点を明らかにする。	Counterfactual Explanation	• ローン審査に落ちた場合に、どうすれば審査を通過するかをCounterfactual Explanationによって確認する(銀行が説明を提供する)。

Unleashing the power of machine learning models in banking through explainable artificialintelligence (XAI)のロールや目的、一部のXAI利用例を参考に作成。

© 2022 The Japan Research Institute, Limited