令和 4 年度 保護観察におけるアセスメントへの AI 導入に関する調査研究業務

# 最終結果報告書

令和 4 年(2022 年) 8 月 31 日

業務受託者
株式会社 AiCAN

本報告書に掲載されている情報は、全て研究業務の受託者が整理・提示した結果および見解であり、事業元である法務省の見解ではない。

# 目次

# 第一章　背景と目的

## 1.1 再犯・再非行と保護観察におけるアセスメント

保護観察において、再犯および再非行(以下、再犯・再非行)の防止は重要な課題である。令和 3 年版の犯罪白書において(法務総合研究所, 2021 年 12 月)、平成 15 年ごろをピークに成人および少年の刑法犯検挙人員は一貫して減少しているものの、令和 2 年の刑法犯検挙人員中の再犯者率は 49.1%、令和 2 年の少年の刑法犯検挙人員中の再非行少年率は 34.7%と報告され、その水準が上昇あるいは維持されている。

再犯・再非行の防止にあたり、犯罪をした人や非行のある少年の処遇は、Risk-Need-Responsivity(RNR)モデル(Bonta & Andrews, 2017)に準拠すること、つまり、再犯リスクの高低に応じた処遇の密度により、犯罪を誘発する要因に焦点を当て、対象者に最も適合する方法で実施することが効果的であるとされている(Bonta & Andrews, 2017; 勝田・羽間, 2020; 羽間・勝田, 2021)。2018 年に、法務省は、Risk-Need-Responsivity モデルを基盤に、保護観察中の人の保険数理的再犯リスク、動的再犯誘発要因及び動的保護要因のアセスメントツールを含む Case Formulation in Probation/Parole (CFP, 以下略記する) を開発した(羽間・勝田, 2021)。これにより、再犯・再非行に関わる動的/静的要因、あるいは保護要因等について多面的な評価がなされ、再犯・再非行リスクの程度に応じた処遇のあり方が検討されてきた。しかし、犯罪や非行が繰り返される背景には様々な要因が関与していると考えられている。正確な再犯・再非行の予測や、その未然防止を実現することは、決して容易ではない。

## 1.2 アセスメント等データの利活用とリスク予測モデリング

諸外国では、Level of Service/Case Management Inventory(Andrews, Bonta & Wormith, 2004)をはじめとして、犯罪をした人の再犯リスクの程度や犯罪を誘発する要因などのアセスメントツールが開発されてきた(羽間・勝田, 2021)。保険数理的アプローチ(actuarial approach)は、こういったアセスメントツールの結果情報を活用し、再犯・再非行リスクの程度を評価する。再犯・再非行の予測力が比較的高い方法であるとされている(Andrews, Bonta, & Wormith, 2006)ものの、蓄積されるアセスメント情報にも限界があることから、必ずしも十分な予測力が得られるとは限らない。CFP 情報を用いた短期的な再犯に対する予測性能を検証した研究(羽間・勝田, 2021)では、広範な測定範囲の情報が得られていないといった制約もあり、研究実施時点において十分と評価されうる予測精度は得られていない。

アセスメント結果に基づく対応が実践される児童虐待対応の領域では、上述の保険数理的アプローチの他、虐待の再発などに関する予測性能の向上を主眼としたアプローチの試行的実践結果などが報告され始めている(e.g. P. Gillingham, 2016)。リスク予測モデリング(predictive risk modeling)と呼ばれ、アセスメントツールの評定情報から予想される結果(e.g. 児童虐待の再度の通告)の出力を、機械学習等を用いたアルゴリズムベース(algorithm-assisted)で行うというものである。具体的には、過去のデータに見られるパターンを機械学習などの技術で捉え、新規に入力した事例情報にそ

れを適用することで、「この事例に見られるパターンの場合には高い確率で再発する」といった予測を実施する。

リスク予測モデリングには、保険数理的アセスメントアプローチには無い複数の利点が存在すると言われている。枚挙すれば、(1) 対象事例では観測されていない情報(欠測)も予測手法の工夫によって過去のパターンから推論することができること、(2) 現在の時代に即した(新しいデータパターンに対応した)パターンが容易に取得できること、(3) 既存のデータを活用することで高い精度での予測が得られる場合があること、(4) 従来のアセスメントツールで行う手続きよりも(自動で実施されるため)一貫性が高いこと、(5) 評定者の専門性に過度に依存せずとも効果的な実装が可能であることなどがそれにあたる(N. Mickelson, T, Laliberte, & K, Piesher, 2017; Marshall & English, 2000, Russell, 2015, Vaithianathan et al., 2012)。このような利点は、従来の保険数理的アプローチで指摘された問題点である予測誤差の大きさ(Baumann et al., 2005)、平均的な傾向を元に構成された項目であるがゆえに個別事例の特徴からくる影響が未考慮となって生じる予測の誤り(Crea, 2010)、合計得点方式等では複雑な現象の生起パターンが捉えきれなかったことによる予測の誤り(CWLA, 2005; Gillingham & Humphreys, 2010)などの課題を改善させることにつながると考えられている(髙岡他, 2020)。

その一方で、リスク予測モデリングに対する批判もある。例えば、(1) 判定結果に過度に依存した機械的なアセスメントにつながるおそれがある、(2) 予測結果が多くの機械学習アルゴリズムにおいて解釈できない、説明されない、(3) 人種や居住地域などによって判定にバイアスが生じる、あるいは、そのような社会的偏見や不平等を強化する(公正なサービスの提供が阻害される)、といったものである(e.g. Blank et al., 2015)。その他にも、(4) 社会問題を個人等の問題に個別化して再定義し、サービス提供の可能性を狭く規定することにつながる(Keddell, 2014)、(5)当初の予定とは異なるデータや予測結果の(益のない)使用が生じる恐れがある、(6)データの欠落や不正確さによる予測の誤りなどの課題があるとされる(e.g. Gillingham, 2016; Braverman et al., 2016)。

リスク予測モデリングへの批判事項に対しては、米国の大学と提携したニュージーランドの研究チームがその対策を具体的に実現しようと試みている。児童虐待予防を扱う当該研究チームによれば、世帯の処遇(措置, placement)や再通告の予測に機械学習を適用する中で、(1) 主に対応の初期段階(コールセンター)における意思決定の主軸を置き換えるものとなるが、サービスの提供にあたっては従来の家族アセスメント実施が前提であって、機械学習の予測結果は「意思決定をよりよくするための参照資源として扱う」ことが明記されており、(2) 結果の説明が必要な場面に対応するため解釈可能性の高い解析手法が採用されており、(3) 人種等によるサービスの偏りが生じないかを検証した上で、例えば再度の児童虐待通告発生の予測性能を中程度以上(All screend in Referrals AUC-ROC = 0.72)で確保したなどの報告がなされている(Vaithianathan, 2017)。Shlonsky et al.(2005)が提示する児童虐待対応におけるアセスメントの将来像に依拠すれば、文脈的・臨床的アセスメントの視点と、各種研究エビデンスを活用する視点に加え、高度な解析を用いた予測モデルによる情報支援を総合的に活用しつつ、対応判断を効果的に構造化してゆくという今後のアセスメントの在り方が提案されている(髙岡他, 2020)。各種の批判に十分に配慮すれば、リスク予測モデリングは、対人援助の専門知識と組み合わせて活用し、十分な思慮に基づくアセスメントとケアに繋がった場合にのみ、その使用が倫理的に健全で社会的に有益なものとなる(Dare, 2015; Keddell, 2014; de Haan, 2014; Shroff, 2017)と整理される。

## 1.3 対人援助関連領域でのリスク予測モデリングの実践に向けた本邦での取組例

本邦では、子どもの生命の危機が懸念される虐待行為や、二次障害の発生等が懸念される重篤な虐待、あるいは児童虐待それ自体の発生などを予測対象とするリスク予測モデリングの実践に向けた取組がいくつか実施されてきている。

令和元年度および令和2年度の子ども・子育て支援推進調査研究事業「児童虐待対応におけるアセスメントの在り方に関する調査研究」では、児童相談所または市区町村への児童虐待通告の比較的初期段階の情報で、重篤な虐待の発生を検知・予測するアセスメントツールの開発が展開された(髙岡他, 2020; 髙岡他, 2021)。具体的には、全国の児童虐待通告事例を対象とする事例調査によって、多面的なアセスメント項目に対する予測的妥当性の基礎評価を実施し、機械学習を用いた場合の各種アウトカムに対する予測性能が試行的に検証された。令和2年度の事業報告では、その調査データに大規模な計画的欠損等が含まれるものの、選抜された20のアセスメント候補項目によって、重度ネグレクト(AUC-PR=0.7116)、重篤な身体的虐待(AUC-PR = 0.6262)、性的虐待(AUC-PR = 0.6533)の併存または将来発生が一定精度で予測できる可能性が示唆されている(髙岡他, 2021)。なお、これらの結果は、子どもの生命の危機や、子どもの健全な発達を阻害する事態の発生を防止するという目的において、十分な精度ではない。子どもの最善の利益を追求するという児童虐待対応分野における最大の理念に照らし、「従来のアセスメントだけでなく、解析技術をも併用して最大限の未然予防・早期発見を実現する」ことに重きが置かれ、「予測結果が対応方針を一意に決定づけることにはならず、有用な参照情報を提供するものである」と利活用上の留意点がガイドされている(髙岡他, 2020)。

母子保健分野では、妊娠届出時の面談や、新生児訪問等事業、乳幼児健康診査の母子保健活動の中で得られる情報から、養育上の不調(一時的な養育困難など)や児童虐待が発生する可能性について見立て、その未然予防と早期発見を実現するためのアセスメントツールが開発され始めている(髙岡他, 2022)。当該研究事業では、「特に支援を必要とする妊産婦・子ども・家庭を把握し、必要な支援を展開する」というハイリスク・アプローチを前提に、機械学習技術を活用範囲に含めたアセスメントツールの初期評価が実施されている。このとき、上述した主要な母子保健活動は、全ての子どもと妊産婦が対象となる。当該研究事業は、アセスメントツールを適用することで生じる偽陽性(アウトカムの発生が予測されたが、実際には発生しない状態)の発生率がシミュレーションされるなど、リスク予測モデリングが及ぼす影響範囲について事前に言及されている点が特長的である。

## 1.4 保護観察領域におけるリスク予測モデリング

Risk-Need-Responsivity(RNR)モデル(Bonta & Andrews, 2017)に準拠する在り方を前提とすれば、各種予防対象の発生リスクを予測的に評価するリスク予測モデリングの枠組みは、保護観察領域において有効に機能する可能性がある。このとき、領域的な性質を鑑みて、第一に想定される予防対象は「再犯・再非行の発生」となるだろう。リスク予測モデリングを保険数理的アプローチの発展形として位置付け、その利点と批判点を踏まえた展開を想定すれば、一例として次のような課題についての検討が必要になると考えられる。

### 1.4.1 前提となる枠組みにおける検討課題の整理(本事業にて想定される範囲での例示)

保護観察領域におけるリスク予測モデリングについて、細部を落とした大きな枠組みを整理するならば、「性別や年齢、罪の内容といった事案の情報を用いて、再犯・再非行の発生を予測し、その結果を各種処遇判断の参考情報として活用する」という流れが想定される。このとき、少なくとも、(1)リスク予測モデルという手段が社会的有益性のある目的に必要な手段となっているか、(2)予測結果の妥当性(predictive validity)が十分あるいは従来のものより優れているか、(3)入力・蓄積されるデータが信頼できるか、(4)構築したリスク予測モデルが処遇機会の公平性を損ねていないか、(5)学習データや、それを活用して得られる予測結果に、差別的な要素が含まれていないか、(6)十分に多面的な情報が学習に用いられているか(学習に使用した情報の範囲が把握されているか)、(7)予測結果をどのような形で判断に援用するのか、(8)リスク予測モデルの各種限界に対する補償・対策がなされているか、(9)設計したリスク予測モデルが各種サービス提供等の業務で実際的に運用可能か、(10)介入判断の論理的根拠や説明可能性をどのように担保するか、(11)データの入力から効果の評価、さらなるデータの蓄積とモデルの更新が自然に循環する仕組みとなっているか、といった点については、事前に十分に検討され、課題が解消されている必要があると考えられる。表 1.1 に、これらの主要な検討課題の一例を整理した。

表 1.1 保護観察領域におけるリスク予測モデリングの前提となる検討課題の例

| 検討観点 | 課題の例 | 検討方法の例 |
|---|---|---|
| 目的との整合性 | ・「再犯・再非行」の発生防止という目的に対して、リスク予測モデリングが妥当な手段となっているか、あるいは、優先的に必要と考えられる手段か。<br><br>・後発した個人データ利活用(リスク予測モデリング)の目的が、元来の個人データ収集の目的に矛盾せず、新たに設定された目的に利用範囲が限定されているか。 | ・現状のアセスメントの課題点と再犯・再非行の予測の困難さを示し、リスク予測モデリングによって部分的な改善が得られることを担保する知見を得る。<br><br>・リスク予測モデリングが、再犯・再非行防止のため、他に真に必要と思われる手段(例. 就労等の各種支援資源の拡充など)を実現するための必要なプロセスの一つとしてあらかじめ位置付けられていることを示す(リスク予測モデリングの結果が、将来的に、目的に整合する、必要な各種施策を提案・展開する上での数値的根拠として利用することなどが事前に計画されているなど)。<br><br>・リスク予測モデリングという形式でのデータ利活用の目的が、元来の個人データ収集の目的に矛盾なく整合しているか検証する。 |
| 予測的妥当性 | ・そもそも、活用するモデルが、適切に再犯・再非行を予測的に評価できているか。 | ・活用するモデルの予測精度指標を算出し、従来の手法(保険数理的アプローチなど)よりも性能が優れていることを示す。<br><br>・人と組織によるアセスメント(Consensus-based |

| | | |
|---|---|---|
| | | approach)の結果と、リスク予測モデルによる予測結果、さらにはこれら両者を組み合わせた総合的な判断結果を比較し、リスク予測モデルを組み入れることで、再犯・再非行が的確に予測できることを示す。 |
| 信頼性・完全性 | ・リスク予測モデルに入力する情報や蓄積データは、利用目的に必要な範囲内で正確、完全かつ最新なものとなっているか。<br><br>・リスク予測モデルに入力する情報や蓄積データが、利用ユーザー間や利用ユーザー内で異なる、または変化する可能性はあるか。 | ・測定時期の古い情報などを参照する場合、対象者の現在の状況を鑑みない予測結果を出力することとなる(対象者の益を損ねる可能性がある)。リスク予測モデルの予測精度が、入力情報の測定時期に依存して異なる可能性を評価する手続きや、予測結果を参照するタイミングと入力情報の測定タイミングが一致するフローの設計などを実施する。<br><br>・アセスメント項目などをリスク予測モデリングに活用する場合は、同一事案や模擬事例などを用いて、評定者内あるいは評定者間で評定値に違いがないかを確認し、ブレの少ない項目等を選抜する(誰が、いつ、どこで評定しても同じ結果が得られる安定した評定項目の研究と選抜)。または、項目の洗練と更新を計画に組み入れる。<br><br>・評定者内および評定者間での評定値の不一致を解消するための、項目定義の明確化や研修などを実施する。 |
| 公平性 | リスク予測モデルによる再犯・再非行の予測結果が、結果的に「処遇機会の公平性を損ねている」といった可能性はないか。 | ・特定の性別や学歴、居住地域の場合に偏って「再犯・再非行の確率が高い」といった予測結果が出力されていないか(性別等の条件で処遇を受ける機会に差が生じる可能性があるか)を定量的に示す。 |
| 非差別性 | リスク予測モデルの予測結果そのものや、結果を活用する際に、対象者に対して、対象者や社会に益のない不要なラベリングが実施されることがないか。 | ・学習するデータが、社会的な偏見等に基づいた結果を含んでいないか。<br><br>・リスク予測モデリングで「再犯・再非行」の発生可能性(リスク)を予測するだけではなく、「対象者がどのようなニーズを有し、それが解消されればどの程度のリスク改善が見込まれるのか」といったニーズの視点をセットにして設計に含め、一側面からのラベリングや、目的外となる誤用の発生を防止する。 |
| 代表性 | 学習されていない範囲の情報は、予測結果を出力する際に直接的に考慮さ | ・あらかじめ、最大限に多面性を与えた学習用データを設計する。 |

| | | |
|---|---|---|
| | れない。再犯・再非行のリスク判断に関連しうる、学習対象範囲外となっている対象者固有の重要な情報はどのように考慮・配慮されるのか。 | ・リスク予測モデルの性質と限界点として捉え、運用上の工夫で補償する仕組みを設計する。 |
| 補償可能性 | リスク予測モデルは的中精度が完全でない限り、「あらかじめ間違った予測が発生することが事前にわかっている」技術であり、「考慮されている情報とされていない情報」が必ず存在し、予測結果が特定の性別等において偏る可能性がある。こういったモデルの限界をどのように補償するか。 | ・リスク予測モデルの性質や限界、学習した情報の範囲について、利用者が必ずそれを理解した上で、判断の補助ツールとして利用するための枠組みを設定する。<br><br>・再犯・再非行のリスクに応じて処遇に濃淡をつけるという、RNR モデルおよびリスク予測モデリングの発想に対して、特定の性別や居住地等による予測結果の不均衡を含め、その不足点を補償する枠組み(対象者による希望を組み入れる機会を確保するなど)を設ける。 |
| 透明性・再現性 | 個人データを用いた開発とその運用について、その利用目的や、当該事業等の実施に伴う責任者に相当するデータ管理者等に関する情報が公開されているか。また、開発と運用に係る手続き等について、その適切性を判断するための情報が記録・整理されているか。 | ・利用する個人データの存在、性質及びその主要な利用目的とともに、責任者等を明示し、容易に閲覧できる形式で公開する。<br><br>・リスク予測モデル構築の手続き、運用上得られた個別の予測結果などが常に再現可能な形式で記録・保存され、モデル構築の手続きや、予測結果、それを参照した判断等の適切性が検証可能な枠組みを設ける。 |
| 活用可能性 | 予測結果をどのような形で判断に活用するのか。そもそも活用できるのか。 | ・保護観察業務のどのようなタイミングで予測結果を出力し、数値をどのように解釈し、処遇決定等の意思決定に際してどのような形でそれを援用し、決定根拠をどのように記述し、介入によってどのような結果が期待されるのか等について、活用方法全体の流れを事前に定める。<br><br>・既存の業務フローに合わせた設計を検討するか、新しい仕組み(リスク予測モデリングを組み入れた業務)に変更する場合には、業務フローの変更を検討する。<br><br>・「処遇の実施等を推奨する」と解釈されるようなモ |

| | | |
|---|---|---|
| | | デルの予測結果がどの程度出現しうるのか。それに対して、実際に処遇を展開できるだけの各種資源は確保されているか。 |
| 解釈可能性・説明可能性 | 内部メカニズムが十分に解釈できない機械学習等によるリスク予測モデルの結果を参照した処遇決定について、その合理性はどのように説明・担保されるのか。 | ・「再犯・再非行」の発生に関する説明原理(理論)を確保し、リスク予測モデルの予測結果を理論に基づく総合的な解釈の一要素として捉える枠組みを設ける。<br><br>・リスク予測モデルの予測結果に応じた処遇決定を試験的(あるいは仮想的)に実施するなどし、「再犯・再非行の発生率が低下する」、「的確な重点処遇が達成される」などの効果を評価し、当該アプローチに基づく判断の有効性を示す。 |
| 循環性・更新可能性 | データ入力・蓄積、モデルの予測、援助等の実践と効果の評価(ユーザー評価を含む)、さらなるデータの蓄積、モデルの更新が、一連の枠組みで自然に循環し、変化に対応できるエコシステムが構築されているか。 | ・リスク予測モデリング全体の設計、実装するシステムの設計などを当該視点から再評価する。<br><br>・実証実験等の試験導入を行い、あらかじめボトルネックとなる箇所を把握し、それを解消する。<br><br>・循環システムの詳細要素に対する定期点検やユーザー意見の収集などを計画に組み入れ、課題抽出と改善が実現される枠組みを設ける。 |
| その他の倫理的妥当性 | ・対象となるリスク予測モデル、設計システム自体、それらを組み入れた実践が、倫理的な妥当性を有しているか。<br><br>・導入・運用の結果、社会的有益性と課題(副作用)を総合的に考慮した評価がなされているか(計画に含まれているか)。 | ・例えば、個人データを活用した再犯・再非行モデルの予測結果は、対象者の個人情報となるか。対象者本人に対する結果のフィードバック等を実施する場合、それが不要なラベリング(あなたは将来再犯・再非行をするだろうと事前に不信を告知されることとなる)などの形で、対象者の益を損ねることに繋がりうる。予測結果の取り扱い、十分な説明と同意など、必要な倫理的配慮と手続きについて、あらかじめ検討する必要がある。 |

### 1.4.2 個別の設計・検討課題(本事業で想定される範囲の例示)

リスク予測モデリングを社会実装する上で、モデル構築等の技術的側面から順を追って整理すれば、次の個別課題を想定することができる(表 1.2)。これらは、表 1.1 に示した前提となる枠組での条件が満たされた上での個別課題として捉えられる。

表1.2 リスク予測モデリングの構築と社会実装に係る個別の検討課題

| 個別観点例 | 概要 | 検討例 |
|---|---|---|
| 予測対象の定義・測定可能性・変容可能性 | ・有益な予測対象は何か(再犯・再非行以外の対象は想定されるか)。<br><br>・予測対象の測定は完全に実現されうるか。<br><br>・リスク予測モデリングの実装後に、その効果として、予測対象の情報は変容するか。 | ・今後の開発・発展可能性を考慮し、個人データ利用の目的に整合し、かつ、有益性が見込まれる他の予測対象項目(予防対象項目等)の記録・測定を行う。<br><br>・「再犯・再非行」といった項目は、事実が検知されなければ測定の完全性が得られない。追跡調査等の実務面での対策や、不完全測定データに対する解析上の対策を行う。<br><br>・リスク予測モデリングの導入(に基づく各種介入の実施)によって、予測対象が影響を受ける場合、予測利用項目と予測対象との関係が適切に学習できなくなるなどの問題が生じる(再犯・再非行のリスクが高いと予測された対象に対して重点的な対応を実施し、その結果、再犯・再非行の発生が防止されたとする。当該データを用いてパターンを学習したリスク予測モデルは、リスク予測モデルに基づく介入の結果を考慮しない限り、本来的にリスクが高いにも拘らず、「再犯・再非行のリスクが低い」という矛盾した予測を出力することになる)。リスク予測モデル導入自体の影響を考慮する設計上の工夫が必要になる。 |
| 予測対象と異なる評価対象の設定 | ・リスク予測モデリングの導入効果として、評価すべき対象があるか。 | ・社会・経済的効果や対象者に対するQOLなど、目的達成の範囲内において、必要となる評価項目を別途設計する。<br><br>・予測出力画面の閲覧回数など、導入効果の背景や課題点を評価するための数値指標(KPI)を設計する。 |
| インプットと学習データの設計 | ・どのようなグループを対象とし、どのような項目や、どのような期間に測定されたデータを学習させるか。<br><br>・予測実行時に、どのよ | ・保護観察の区分や、犯罪の種別など、どのような単位で予測モデルを構築するか。<br><br>・どのような項目の情報を収集し(どのような項目を収集しないか)、どのような形式で学習に使用するか(エンコーディング、次元削減など)。またそれは、予測実行時の入力形式と合致するか。 |

| | | |
|---|---|---|
| | うな項目を利用するか(利用可能か)。 | |
| モデルの選択と評価 | ・どのようなモデルや、モデルの組み合わせを予測に使用するか。<br><br>・性能評価のための検証データは適切か。 | ・相対的に解釈性に優れるモデルや、処理が軽量のモデル、予測性能を重視するモデルなど、目的と制約に応じたモデルを選択する。<br><br>・予測対象の発生メカニズムが時期に依存して変化する可能性があるなどの場合、予測性能を評価するための検証データを「最新年の記録データ」とするなどの手段が想定される。 目的に応じた評価手法や指標を選択する。 |
| アウトプットの設計 | ・予測結果をどのような形式で出力・表示するか。 | ・「判定結果」「リスク水準」「スコア」「確率」など、どのような形式で予測結果をアウトプットするか。感度と特異度などのバランスなど、目的に応じた出力となっているか。<br><br>・個別の予測結果に対して「モデルがどのようにその予測結果を出力したのか」に関する要約情報を出力するか否か(不用意な解釈情報の提示は誤った因果律を錯覚するなどの誤解を招く可能性がある)。 |
| 蓄積データ設計 | ・差別や偏見などの社会的バイアスが排除されたデータの蓄積ができているか。<br><br>・信頼性の担保された整然なデータの蓄積がなされているか。 | ・社会的な偏見に基づく評定結果等が蓄積される設計となっていないか。<br><br>・評定者によって過度に異なる評定結果が得られる情報項目が含まれていないか。<br><br>・データの欠落や不備の発生を招く、入力負担などが軽減・排除されているか。<br><br>・項目追加などのデータ更新について、既存のモデルを運用しながら、新たな項目データが自然に蓄積されるような仕組みが得られているか。 |
| モデル更新の設計 | ・時間経過に伴う環境の変化や新規データの蓄積による、モデル更新が事前に設計されているか。 | ・定期的なモデルの評価(系統的な予測の誤りなどが発生し始めていないかなど)と、性能向上や各種事態に対応するためのリスク予測モデルの更新作業が、あらかじめ計画されているか。 |
| 実装媒体とシステムの検討 | ・業務フロー等に沿ったリスク予測モデリングの | ・リスク予測モデルにデータを入力する状況と、学習に使用するデータが得られた状況が一致する仕組 |

| | | |
|---|---|---|
| | 循環システムが成立するための、ユーザー端末やシステムサーバ、データベースなど、バックエンドからフロントエンドまでのシステムが一貫して設計されているか。 | みとなっているか。<br><br>・セキュリティや情報保全のための対策は十分に施されているか。<br><br>・ユーザーが予測結果を利用する場面、状況、当該場面で必要とする情報が提示されるのにふさわしい端末・媒体で実装されているか。<br><br>・業務の基幹となるシステムと結合されているなど、二重業務等の発生を防止する設計となっているか。 |
| 利便性評価 | ・ユーザー視点から、システムの利便性が確認されているか。 | ・データの入力や予測結果を参照する画面が、ユーザーの視点から設計され、必要かつ十分な情報が提示されているか。<br><br>・システムが動作するか(エラーがないか)、不便ではないか、便利であるか、ユーザーや対象者に良い体験を与えているか、なくてはならないものであるかなど、ユーザーと、ユーザーを介して益を得る対象者の視点から、システムの利便性を評価する。 |
| 導入整備 | ・設計意図に沿ったリスク予測モデルの使用が展開されるための事前準備がなされているか。 | ・十分な動作に必要な端末や通信環境などの情報インフラが整備されているか。<br><br>・運用に必要な各種体制の整備がなされているか。<br><br>・マニュアルやガイドブックの整備、リスク予測モデルの適切な運用に必要な知識・実践に関する研修の実施など、利活用と定着に係る取組が継続的になされているか。 |
| 点検と効果評価 | ・リスク予測モデルが目的や意図と異なる動作をしていないか。<br><br>・リスク予測モデルの導入によって、目的とする効果が得られているか。 | ・定期的に、入力情報の変遷(入力ドリフト)や、アウトプット情報の異常値出力などを点検する枠組みが設けられているか。<br><br>・意図と異なる予測が出力される事態が発生する場合に、動作を停止する枠組みが与えられているか。<br><br>・「導入の効果」を評価する枠組みの事前設計がなされているか(効果評価に必要な情報が、データとして蓄積される仕組みになっているか) |

## 1.5 事業の目的と検証課題

前節の表 1.1 および表 1.2 に一例を整理した通り、保護観察領域に固有と思われる観点も含め、リスク予測モデリングを社会実装するために検討すべき課題は多い。本事業では、リスク予測モデリングの構築に向けて、既存のデータを用いた試験的解析を実施する。機械学習モデルによる再犯・再非行を予測対象とする解析を実施することで、現状データでの予測的妥当性(予測精度)を把握する。また、本事業では、今後主要な利用項目になると想定される CFP アセスメントデータに関して、その基礎的な特徴把握を行う。そして、これらの作業を通じて得られた知見や課題点を含めて、今後検討が必要と想定される事項を、ロードマップ案の形式で整理する。事業目的と対応する個別検証課題を表 1.3 に整理する。

表 1.3 本事業の目的と個別の検証課題

| 該当章 | 目的 | 作業区分 | 本事業での部分検証課題 | 補足と備考 |
|---|---|---|---|---|
| 第二章 | 保護観察における 5 年以内の再犯・再非行のリスク予測モデルについて、その実装のための基礎的な検証を行う。 | 5 年以内の再犯・再非行を予測対象とし、既存の基礎データを用いた場合の予測的妥当性を評価 | ・機械学習モデルを用いた予測精度の評価<br>・保護観察の号種で分けたモデルの試験構築と予測精度の評価<br>・経年でデータを分割して予測精度を評価する(出力に与える入力情報の経年変化の基礎評価) | 採用モデルの元で予測に貢献した項目を把握する |
| 第三章 | 保護観察における比較的短期での再犯・再非行の予測を通じて、CFP データの基礎的特徴を把握する。 | 蓄積された CFP データ(テキスト情報を含む)の特徴把握 | ・CFP 要因区分ごとの単語頻度等情報を集計し、再犯・再非行との関係情報を抽出する<br>・CFP データの情報が、比較的短期での再犯・再非行の予測に貢献する度合いを機械学習モデルで評価する | 基礎データの集計等は、犯罪白書(e.g. 法務総合研究所, 2021)に示された内容との重複が見込まれるため、CFP アセスメント情報を対象とした。 |
| 第四章<br>・<br>第五章 | 本事業の解析結果を踏まえ(第四章)、本事業内で想定される今後の検討課題を整理する(第五章)。 | ロードマップの作成 | ・保護観察領域でのリスク予測モデリング(AI-CFP)の構築に求められる要素を整理 | |

# 第二章　事案基礎データを用いた 5 年以内再犯・再非行の予測

## 2.1 解析の目的と検証課題

本章では、保護観察における 5 年以内の再犯・再非行のリスク予測モデルについて、その実装のための基礎的な検証を行う。具体的には、既存の基礎データ(事件管理システムのデータ)を用いた場合の予測的妥当性に焦点を当て、(1)保護観察の号種別での試験的モデル構築と予測精度の評価、(2)特定期間でデータを分割した上での予測精度の評価を行う(出力に与える入力情報の経年変化の基礎評価)。

## 2.2 方法（データ抽出）

### 2.2.1 データ源からのデータ抽出

保護観察対象者に関連するデータを結合したテーブルである「調査研究用データ」は、本事業調査研究用に事前結合された、対象者の事案に関連する各種基礎情報が格納されたテーブルである。当該テーブルは、事件管理システムに含まれる複数テーブルの情報を元に、事業依頼元管理のもと、事業依頼元の担当者によって、2022 年 5 月 10 日に結合処理と合わせて抽出・作成された。解析対象レコードの抽出前における当該データセットには、462 項目に関する 432,721 レコードが含まれた。

### 2.2.2 解析対象レコードの抽出

調査研究用データから解析対象レコードを抽出するにあたっては、(1)保護観察の号種が 5 であるレコード、(2)同一対象者において同一保護観察期間開始日が付与された重複レコードのうち欠損率の高いレコード、(3)5 年以内の再犯・再非行ラベルを付与したのちにデータ抽出日(2022 年 5 月 10 日)から遡って 5 年となる 2017 年 5 月 10 日以降に保護観察期間開始日が付与されたレコードを追跡期間確保のために除外した。

項目の抽出にあたっては、(1)全てのレコードが欠損している項目または 99%以上を目安として著しい欠損が含まれる項目を除外し、(2)保護観察開始前の各種情報、保護観察開始時、保護観察期間中に取得可能な項目を組み入れ対象とし、(3)予測対象と明らかに無関係であると解析者によって判断された日時や ID 等の項目を除外する、という条件を設定した。

### 2.2.3 データ区分の定義と分割

2.3.4 章では各項目がアウトカムの予測にどの程度貢献しているかを評価する指標について述べる。この指標は、少なくとも「保護観察の号種」によって解釈のあり方が異なる可能性がある。よって、本解析では、号種 1(いわゆる一号観察)から号種 4 までのそれぞれについて解析を実施することを前提に、号種でのサブグループ設定を適宜行う。

また、経年でのデータの質的な相違が予測結果へ与える影響を評価するために、三号観察と四号観察の対象データについては、それぞれの号種ごとに、2008 年 5 月から 2016 年 5 月まで(以後、前期区分と呼ぶ)の対象データと 2016 年 6 月(一部執行猶予制度の開始)から 2017 年 5 月まで(以後、後期区

分と呼ぶ)の対象データに分けて一部の解析を行う。なお、期間によるデータの分割は 2.3.5 章とその結果に対応する 2.5.5 章における解析のみで実施する。

## 2.3 方法(機械学習)

### 2.3.1 アウトカムの設定

本事業では「保護観察開始日から５年以内に再度、犯罪又は非行をして保護観察所に係属すること」を再係属と定義し、これをアウトカムとする。本事業で用いたデータにおける再係属率等の基本統計量は 2.5.2 章で報告する。

### 2.3.2 機械学習モデル

保護観察開始時(一部処遇の情報も含む)のアセスメント情報から再係属の有無を予測するための手法として、Light Gradient Boosting Machine(LightGBM)を採用した。LightGBM とは、勾配ブーストと呼ばれる最適化アルゴリズムを用いた機械学習手法の一種であり、対象が観測されるパターンを条件分岐の形式で捉えるモデルである(Guolin Ke el al., 2017)。このモデルの特徴は予測性能が比較的高い点、計算時間が比較的短い点、メモリ消費が抑えられる点である。なお、LightGBM と同様の手法に eXtreme Gradient Boosting(XGBoost)があるが、LightGBM は XGBoost をもとに考案されたモデルであり、計算時間やメモリ消費の観点から改良が行われた手法である。

LightGBM を実装する際は、統計解析環境 R(version 4.0.3, R Core Team, 2022)と、lightgbm パッケージ(version 3.3.2, Shi et al., 2022)を用いた。また、学習(最適化)の目的関数には、対数損失(logloss)を採用した。

機械学習を用いた予測を行う際は、手元にあるデータセットを(1)学習データ(training)、(2)ハイパーパラメータ調整用のデータ(validation)、(3)性能評価用の検証データ(test)に分割する。本章の解析ではこれらをそれぞれ 55%、25%、20%の割合で無作為に分割して作成した。

ハイパーパラメータの調整は、learning_rate(学習率)、num_iterations(使用する決定木の最大数)、num_leaves(各決定木における葉ノード数の上限)を対象とした。それぞれ、learning_rate={0.5, 0.1, 0.01}、num_iterations={100, 1000}、num_leaves={15, 31, 60, 120} を候補の値とし、ハイパーパラメータ調整用のデータ(validation)での AUC-ROC が最大になるように調整を行った。

learning_rate と num_iterations はグリッドサーチによる交差検証法(3-fold cross validation)を実施し、num_leaves は他のハイパーパラメータとは独立に交差検証法(3-fold cross validation)を実施した。さらに、2.5.2 章で述べるように学習データの再係属率は不均衡であり、これによる精度の低下を和らげるために、bagging_freq(各決定木を学習するためのデータを無作為に抽出する頻度)を 1 とし、neg_bagging_fraction を学習データにおける再係属率/ (1-再係属率)に設定した。これにより、モデルの学習の際に、再係属に該当する事案と非該当の事案が均等にサンプルされる。上記で述べていない LightGBM のハイパーパラメータについては、lightgbm パッケージのデフォルト値を設定した。

最後に、本章の解析では再現性を担保するために、乱数のシード値を固定して解析を行った。

### 2.3.3 性能評価方法

機械学習をはじめとするモデルの予測性能を評価する際には、様々な指標が目的に応じて使い分けられる。全体的な性能を評価する指標や、ある部分に着目した指標などが存在する。本章の解析では、総合的な指標と部分的指標の両方を使用・報告し、結果を多面的に評価する。

総合的な指標としては、PR 曲線下面積(Area Under the Precision-Recall Curve: AUC-PR)を主要な指標として使用する。AUC-PR は、横軸に感度(Recall)、縦軸に陽性的中率(Precision)を取り、機械学習が出力する予測スコアの判別閾値を変化させた時に現れる曲線(PR 曲線)で区切られた範囲下の面積である。アウトカムの該当率(25%の場合は 0.25)から 1 までの値を取り、最小値からの値の向上が大きいほど高い予測性能を有することを示す。本章の解析におけるアウトカムである再係属の該当率は、50%を下回り(2.5.2 章参照)、クラス不均衡と呼ばれる状態にある。AUC-PR はクラス不均衡における性能評価の際に、結果の誤解を招くリスクが低減されることなどを理由に、その利用が推奨されている(Sofaer et al., 2019)。ただし、AUC-PR の数値は、大小の比較を除いて、その値の持つ意味を解釈することが難しい。そこで、本解析では、補足的に Accuracy(精度)を総合的な評価指標として報告する。Accuracy とは、機械学習のアウトカムに対する該当/非該当の予測と、実際のアウトカムの該当/非該当が一致していた割合(正解率)を示す指標である。これは、0 から 1 までの値をとり、クラス不均衡の場合に、大きな値に偏りやすいという性質をもつが、解釈が容易であるという利点をもつ。

上記で述べた AUC-PR と Accuracy に加え、保健・医学分野で頻繁に使用される指標である ROC 曲線下面積(Area Under the Curve of Receiver Operating Characteristic: AUC-ROC)もあわせて報告する。AUC-ROC とは、縦軸を感度、横軸を 1-特異度とし、閾値を変化させた時に現れる曲線下の面積である。0 から 1 までの値を取り、値が大きいほど予測性能が高いことを示す(値が 0.5 の時はランダムな予測と同等であり予測性能がほとんどないことを意味する)。ただし、AUC-PR とは異なり、正例の少ないクラス不均衡の場合には、AUC-ROC の数値から得られる印象よりも陽性的中率(後述)が低くなるという「誤解を招きやすい」性質をもつことから、解釈の際には多面的に複数の指標を参照されたい。

総合的な指標とは別に、部分的な評価指標として、閾値を任意に定めた場合の(1)Recall(感度)、(2)Specificity(特異度)、(3)Precision(陽性的中率)、(4)Negative Predictive Value(陰性的中率)も評価する。これらの指標はそれぞれ、(1)Recall(感度):実際のアウトカムに該当の事案のうち正しく該当であると予測できた割合、(2)Specificity(特異度):実際のアウトカムに非該当の事案のうち正しく非該当であると予測できた割合、(3)Precision(陽性的中率):アウトカムに該当であると予測された事案のうち実際に該当である割合、(4)Negative Predictive Value(陰性的中率):アウトカムに非該当であると予測された事案のうち実際に非該当である割合となる。以降、それぞれの指標を日本語表記する際にはそれぞれを(1)感度、(2)特異度、(3)陽性的中率、(4)陰性的中率と表記する。

### 2.3.4 項目の評価方法(Shapley Additive Explanations)

2.3.1 章では、保護観察開始時(一部処遇の情報も含む)のアセスメント項目から再係属の有無(アウトカム)を予測する機械学習モデルの構築方法について述べた。これに加え、本章の解析では、アウトカムの予測にどの程度貢献しているかという観点から各アセスメント項目を評価する。その指標として、学習済みの機械学習モデルに対する SHAP(SHapley Additive exPlanations) と呼ばれる値を算出する(Scott and Lee, 2017)。SHAP は、ある項目がアウトカムの該当予測に貢献した際に正の値を取

り、非該当の予測に貢献した際は負の値を取る。また、項目の貢献度が大きいほど対応する SHAP の絶対値も大きくなる。SHAP は、各事案(サンプル)の各項目についてそれぞれ計算される。各項目の SHAP の絶対値を、与えられた全事案(サンプル)について平均したものは Global SHAP と呼ばれ、各項目の包括的な貢献度を表す。本章の解析では、すべてのアセスメント項目について Global SHAP の値を評価し、その値を項目間で比較する。なお、SHAP を用いた解析には限界が存在する。 SHAP はあくまで、予測モデルをブラックボックスとして扱い、その入力値と出力値だけに着目して予測的貢献度を評価する手法である。よって、SHAP は貢献度の評価基準として確立された指標ではあるが、なぜそのような予測になったかを演繹的に説明することはできない。

### 2.3.5 対象期間別の精度評価

本章における上記の手続きでは、保護観察開始日が 2008 年 5 月から 2017 年 5 月（以後、全期間データと呼ぶ）までの比較的長い期間のデータを使用して機械学習モデルを学習した。しかし、刑法の改正等により、経年でのデータの質的な相違が潜在している可能性がある。これに起因した予測結果への影響を概観するため、三号観察と四号観察の対象データについては、それぞれの号種ごとに、2008 年 5 月から 2016 年 5 月まで(以後、前期区分と呼ぶ)の対象データと 2016 年 6 月(一部執行猶予制度の開始)から 2017 年 5 月まで(以後、後期区分と呼ぶ)の対象データに分けて精度評価を実施した。

全期間の学習データで学習を行った機械学習モデルに対して、検証データ（学習に使っていないデータ）を前期区分データと後期区分データに分けて精度評価を行い、それらを比較する。もし、対象期間区分ごとにデータの質的な相違が潜在する場合、精度評価の結果に違いが観測されることが予想される。

## 2.5 機械学習を用いた 5 年以内再犯予測の結果

### 2.5.1 データ抽出の結果

2.2.2 章で述べた解析対象レコード抽出手順に則り、(1)保護観察の号種が 5 であるレコード：4 件、(2)同一対象者において同一保護観察開始日が付与された重複レコードのうち欠損率の高いレコード：62 件、(3)5 年以内の再犯・再非行ラベルを付与したのちにデータ抽出日(2022 年 5 月 10 日)から遡って 5 年となる 2017 年 5 月 10 日以降に保護観察期間開始日が付与されたレコード：120,826 件を除外した。その結果、解析対象レコードとして 311,829 件が抽出された。

項目の抽出についても、2.2.2 章で述べた手順に則ることで、最終的に 53 の項目が解析対象として抽出された。抽出した項目については、補足資料図 S2.2 を参照されたい。

### 2.5.2 主要変数の基本統計量

本解析のアウトカムである再係属に関する基本統計量を報告する。データ抽出処理後の全事案データにおける、再係属有無の件数と割合を図 2.1 にまとめた。全事案の 28.8%が再係属に該当し、該当率と非該当率が不均衡(クラス不均衡)であることが見られる。また、一号観察から四号観察の事案別に同様の解析を行った結果が図 2.2 である。再係属有無の該当率は号種別で見ても概ね変わらない(24%から 33%)。どの号種でもクラス不均衡の状態にあることが分かる。

2.3.5 章で述べたように、三号観察と四号観察のサブグループでは、期間によるデータ分割を行う。号種別・期間別でデータを分割した際の、再係属有無の件数と割合を図 2.3 に示す。ここでも、全てのデータ区分においてクラス不均衡が見られる。なお、同一号種区分において、異なる期間区分間での再係属の該当率は差異が小さい。

次に、再係属に該当する事案における犯歴回数ごとの件数とその割合(再係属に該当する全事案が母数)を図 2.4 に示す。再係属に該当する事案の約 2 割は犯歴が 2 回以上であることが分かる。これはデータの中で、同じ対象者のレコードが重複で何件含まれるかに関わる指標である。

同様の解析を号種別にまとめたのが図 2.5 である。ここでは号種による分布の違いが見られる。このことから、号種別でサブグループの設定を行った際には、各グループ間でデータの質的な相違が潜在する可能性が示唆された。

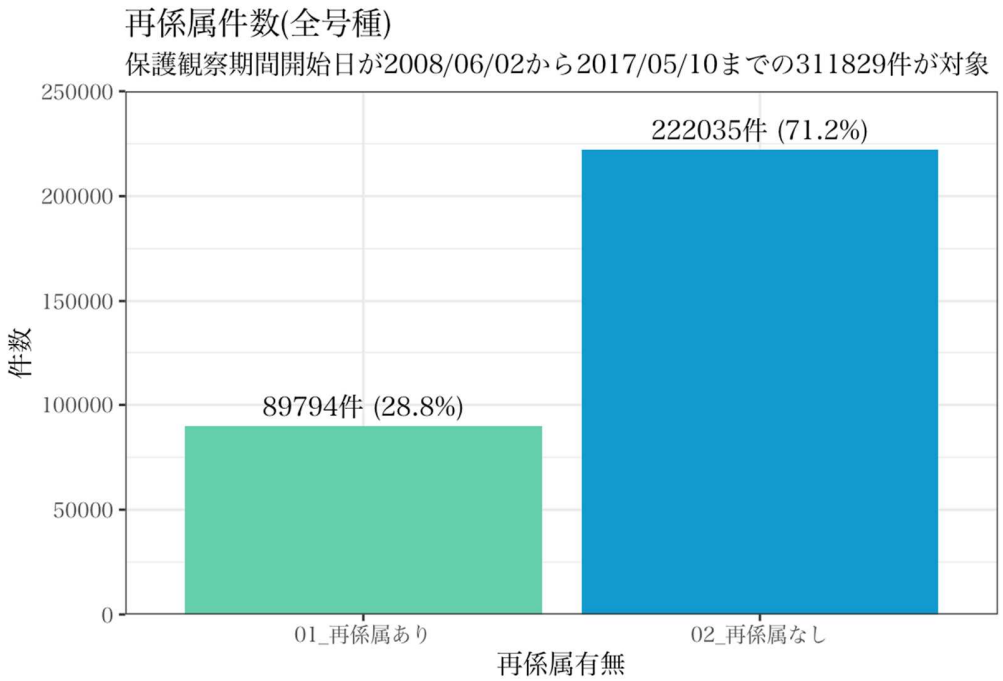



図 2.1 再係属件数(全号種)

解析対象レコードの 311,829 件が対象

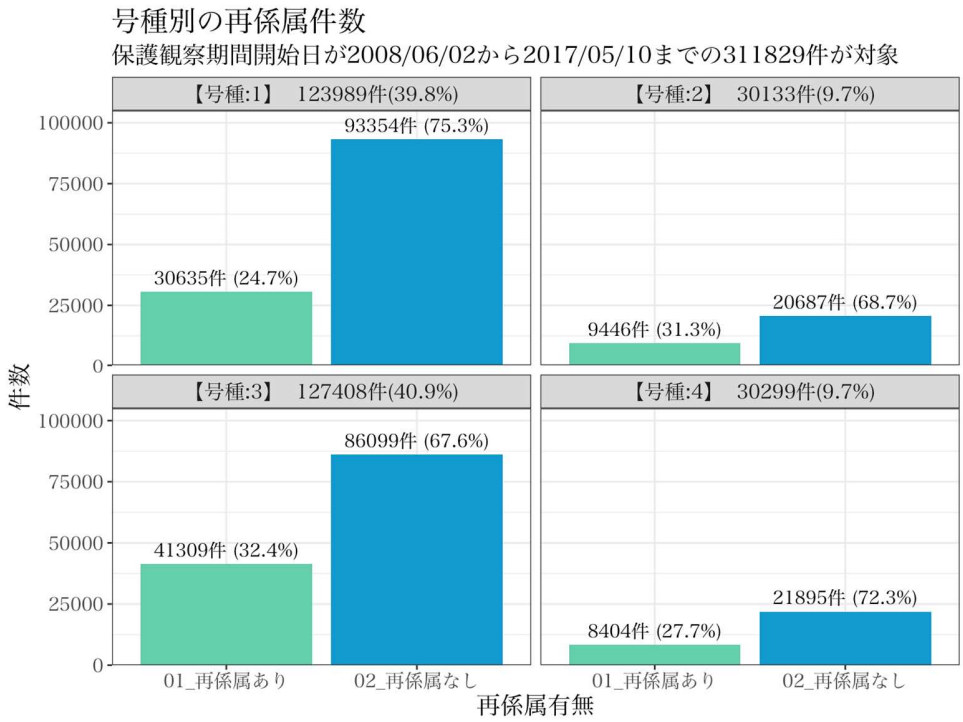



図 2.2 再係属件数（号種別）

解析対象のレコード数はそれぞれ一号観察：123,989 件

二号観察：30,133 件、三号観察：127,408 件、四号観察：30,299 件

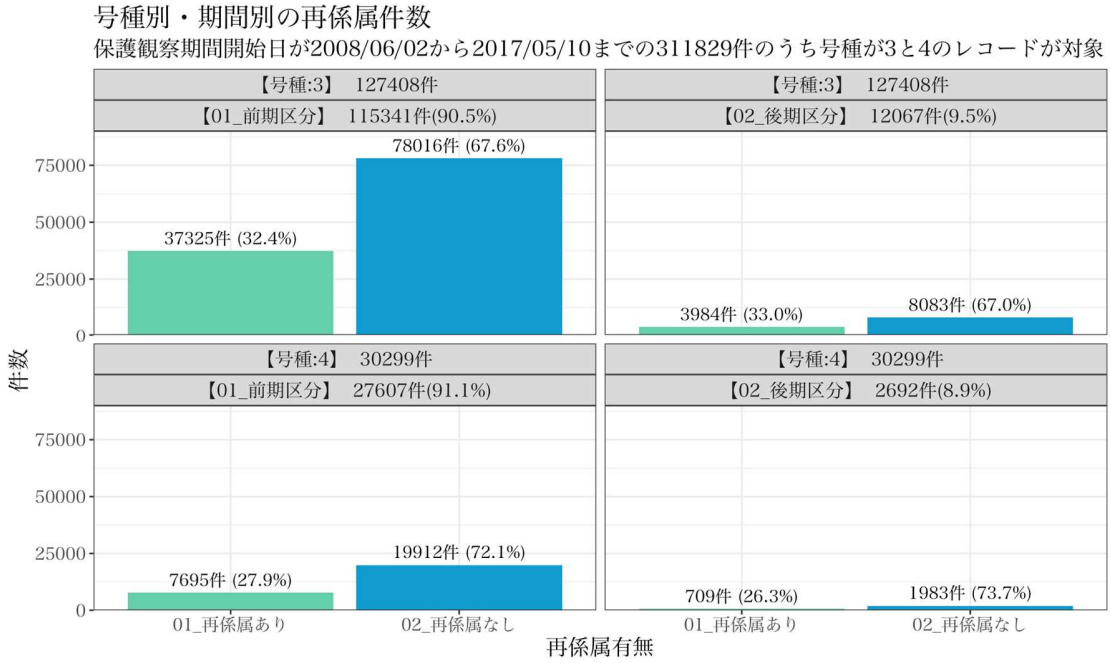



図 2.3 再係属件数（号種別・期間別）

解析対象のレコード数は号種別にそれぞれ、三号観察：127,408 件、四号観察：30,299 件

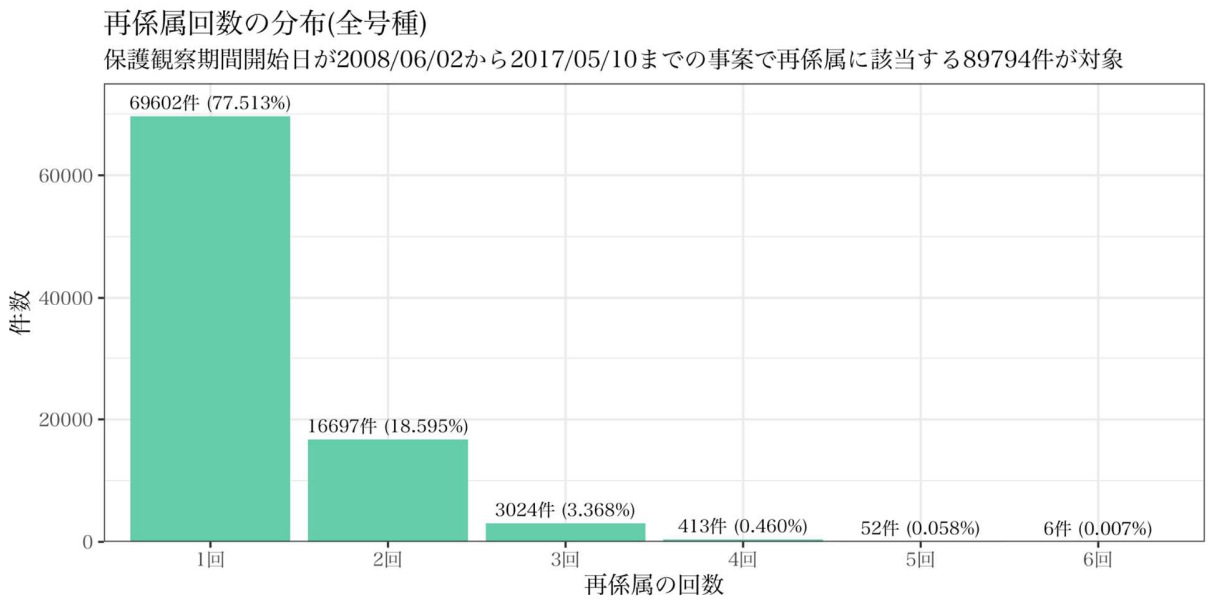



図 2.4 再係属の回数別発生頻度（全号種）

再係属に該当する 89,794 件が対象

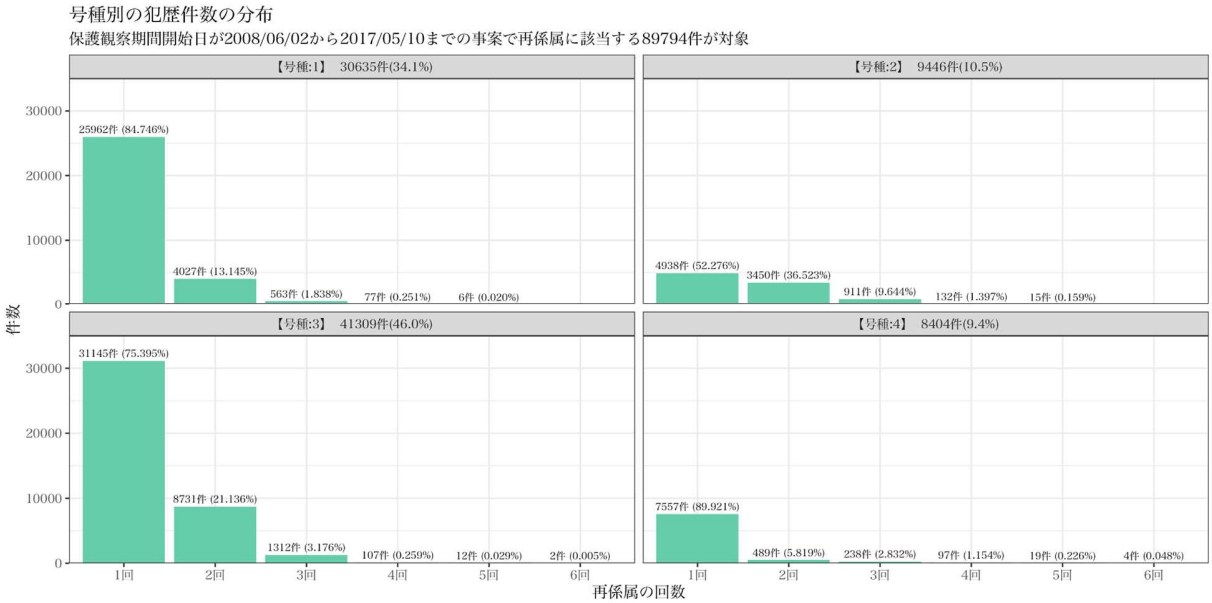



図 2.5 再係属の回数別発生頻度（号種別）

再係属に該当する 89,794 件が対象。号種別の解析対象レコード数はそれぞれ一号観察：30,635 件、二号観察：9,446 件、三号観察：41,309 件、四号観察：8,404 件。

### 2.5.3 機械学習による予測モデルの精度

号種別に分割したデータと全データ(全事案) のそれぞれに対して、2.3.2 章で述べた手順で機械学習モデルを構築した。ハイパーパラメータ調整用のデータ(validation)に対して目的関数を最小化できるハイパーパラメータの組み合わせは表 2.1 のようになった。このハイパーパラメータのもとでモデルを構築し、2.3.3 章で述べた各種精度評価指標を性能評価用の検証データで算出した結果が表 2.2 である。性能指標に関する具体的な解釈例は、補足資料 S2 の記載例を確認されたい。各結果に対応する機械学習の予測値の分布、PR 曲線、ROC 曲線は補足資料図 S2.1 を参照されたい。なお、データの母集団は号種ごとに異なる可能性があるため、AUC-ROC や AUC-PR の直接的な比較はできないことに留意されたい。また、それぞれの号種において、データのサンプルサイズが異なり、一般的にはサンプルサイズが十分である時に、より高い精度が出ることにも注意されたい。今回、全データ(全事案)のサンプルは 311,829 件であり、号種ごとにサブグループ化することで、一号観察：123,989 件、二号観察：30,133 件、三号観察：127,408 件、四号観察：30,299 件であった。

表 2.1 号種別にデータを分割し各データで機械学習モデルを構築した際の最良ハイパーパラメータの組

号種:All は全データ(全事案) に対して機械学習モデルを構築した際の解析結果を示す。

| 号種 | learning_rate | num_iterations | num_leaves |
|---|---|---|---|
| All | 0.01 | 852 | 31 |
| 1 | 0.01 | 604 | 15 |
| 2 | 0.01 | 364 | 60 |
| 3 | 0.01 | 490 | 120 |
| 4 | 0.10 | 47 | 31 |

表 2.2 号種別にデータを分割し各データで機械学習モデルを構築した際の各種精度評価指標

号種:All は全データ(全事案) に対して機械学習モデルを構築した際の解析結果を示す。Threshold(閾値) は、機械学習の予測スコア(0 から 1 の連続値)を該当/非該当(0 か 1 の二値)に変換する際の閾値を表す。

| 号種 | Threshold | AUCROC | AUCPR | Accuracy | Precision | Recall | Specificity | NegPredValue |
|---|---|---|---|---|---|---|---|---|
| All | 0.480 | 0.747 | 0.521 | 0.646 | 0.437 | 0.768 | 0.596 | 0.863 |
| 1 | 0.484 | 0.738 | 0.450 | 0.647 | 0.385 | 0.741 | 0.617 | 0.881 |
| 2 | 0.482 | 0.717 | 0.547 | 0.657 | 0.480 | 0.672 | 0.649 | 0.805 |
| 3 | 0.472 | 0.760 | 0.563 | 0.652 | 0.475 | 0.796 | 0.584 | 0.858 |
| 4 | 0.514 | 0.713 | 0.463 | 0.650 | 0.418 | 0.679 | 0.638 | 0.838 |

### 2.5.4 項目の評価

2.5.3 章で構築したモデルに対して、2.3.4 章で述べたように、すべてのアセスメント項目について Global SHAP の値を評価し、その値を項目間で比較した。その結果、それぞれの号種別サブグループにおける上位 3 項目は、一号観察［1 位：保護観察期間開始日_年齢、2 位：X18_罪名.非行名 1、3 位：X14_教育程度コード]、二号観察 ［1 位：保護観察期間開始日_年齢、2 位：X18_罪名.非行名 1、3 位：X14_矯正入所回数]、三号観察 ［1 位：X18_罪名.非行名 1、2 位：X12_入所.入院度数コード、3 位：X14_保護観察回数コード]、四号観察： ［1 位：X18_罪名.非行名 1、2 位：X14_刑事処分歴コード、3 位：保護観察期間開始日_年齢］ となった。また、全事案(全号種)に対するモデルの上位 3 項目は、［1 位：X18_罪名.非行名 1、2 位：保護観察期間開始日_年齢、3 位：X14_X14_保護観察回数コード]であった。どの号種サブグループにおいても、「X18_罪名.非行名 1」は比較的上位に挙がることが観測された。全項目の順位については補足資料図 S2.2 を参照されたい。

補足資料図 S2.2 から分かるように、どの号種においても、保護観察期間中の情報である処遇プログラムに関する項目は再係属に対する予測的貢献度(Global SHAP の値)が比較的小さい。文献(髙岡他, 2022)で実践されているように、予測的貢献度の値が比較的小さい項目を項目群から除外して再度機械学習モデルを構築した場合、その予測精度は除外前の精度と大きく変わらない。実際の現場に機械学習モデルを実装する際は、処遇プログラムに関する項目に限らず、予測的貢献度の低い項目を除外することで、現場の負担感を減らせる可能性がある。

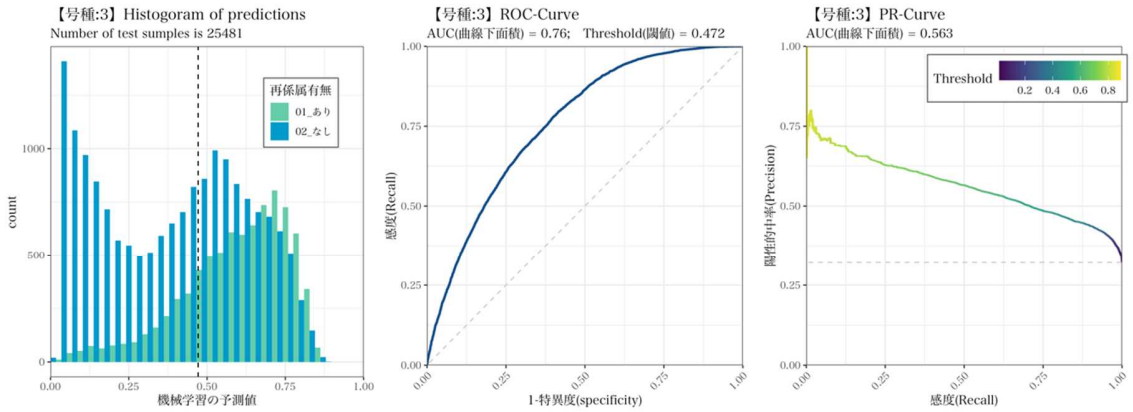
### 2.5.5 対象期間別の精度評価結果

2.3.5 章で述べたように、三号観察と四号観察のサブグループそれぞれに対して、全期間のデータ(学習データの全事案)を使用して機械学習モデルを学習した後、前期区分データ(検証用データの一部)、後期区分データ(検証用データの一部)、全期間データ(検証用データの全事案)のそれぞれで精度評価を実施した。機械学習の予測値の分布、PR 曲線、ROC 曲線を図 2.6(三号観察の結果)と図 2.7(四号観察の結果)に示す。三号観察、四号観察共に、後期区分データにおける AUC-PR の値が全期間データ、前期区分データのそれを上回っていることが見られる。ここから、後期区分データに対しては、再係属の予測がより容易であることが示唆される。期間によるデータ区分間で使用項目は同一であるため、項目の該当パターンの相違が本結果の要因となっている可能性がある。ただし、制度の改正等により、アウトカム(再係属)の定義や該当パターンが変化し、それが予測精度に影響を及ぼしている可能性もある。

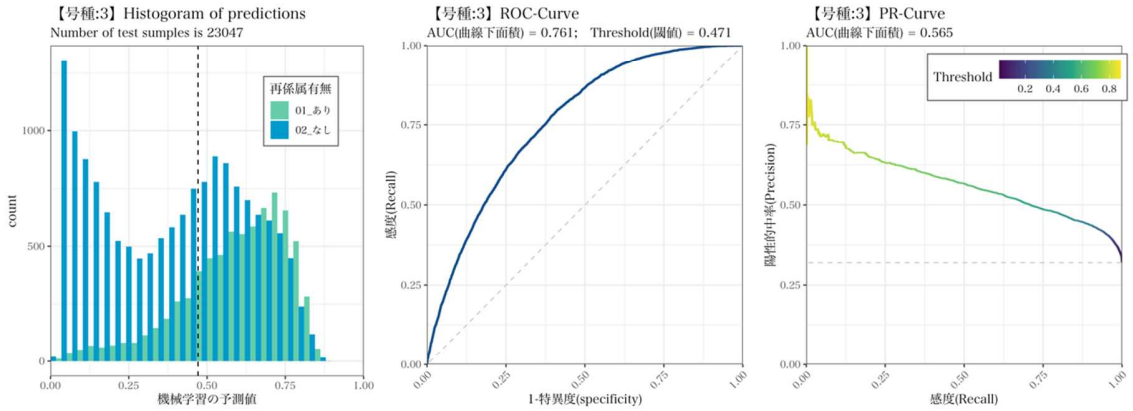
なお、同一号種区分であっても、期間区分の異なるデータ間では母集団の特性が異なり、AUC-PR の比較が必ずしも正当化されないことに留意されたい。また、データ区分ごとのサンプル数の差も結果に影響を与える可能性がある。

上記の結果を踏まえ、後期区分だけの学習データを使用して機械学習モデルを構築し、後期区分の検証用データで精度評価を行った。その結果、AUC-PR や AUC-ROC などの総合的な精度評価指標は、全期間データで学習を行った際の結果に比べ低下した(補足資料図 S2.3,図 S2.4)。学習データのサンプルサイズが小さくなったことが一つの要因であると考えられる。機械学習モデルを実装し、活用する際には、データの経年変化を考慮した期間区分に加え、確保可能なデータのサンプルサイズなども考慮し、総合的な評価の上でモデルを構築する必要があることが示唆された。

【全期間】




【前期区分：2008年5月から2016年5月】




【後期区分：2016年6月（一部執行猶予制度の開始）から2017年5月】

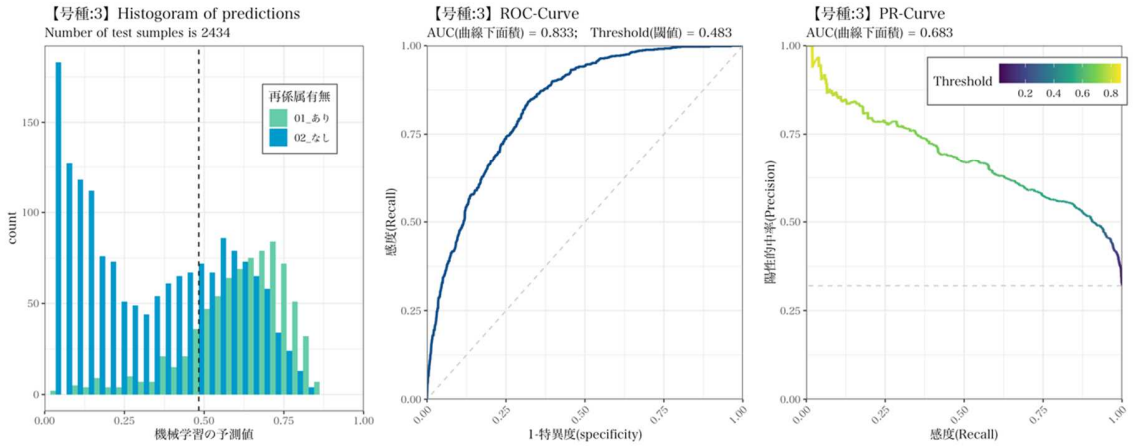
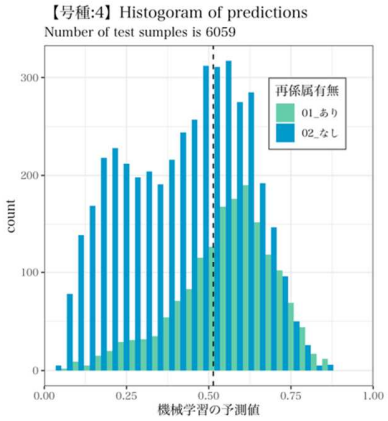
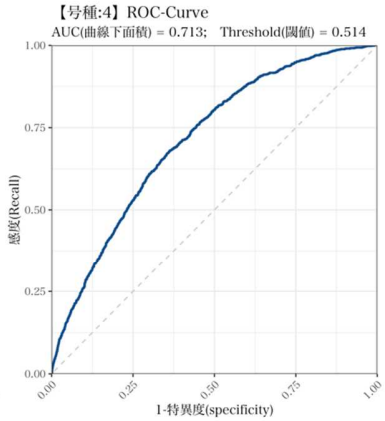


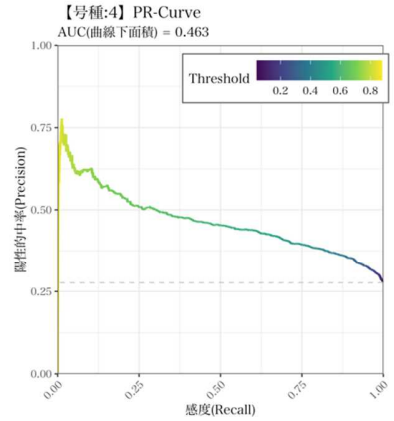

図 2.6 三号観察に該当する学習データの全事案(全期間)で学習を行った機械学習モデルに対する、前期区分データ(検証用データの一部)、後期区分データ(検証用データの一部)、全期間データ(検証用データの全

事案)における予測結果。異なる列は左から順に、機械学習の予測値分布、ROC 曲線、PR 曲線を示す。異なる行は上から順に、前期区分データ、後期区分データ、全期間データに対する結果を示す。

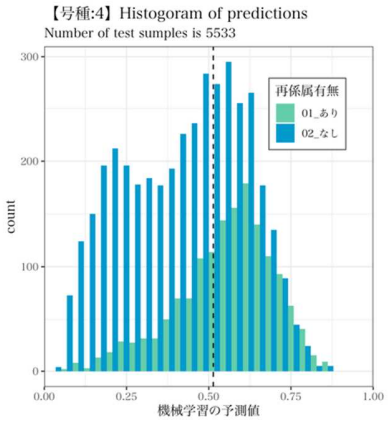
【全期間】

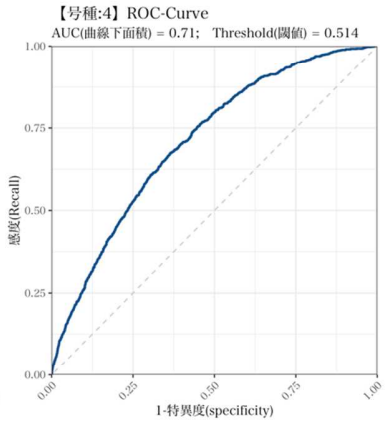


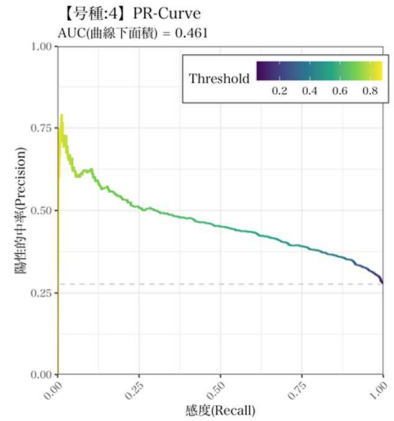



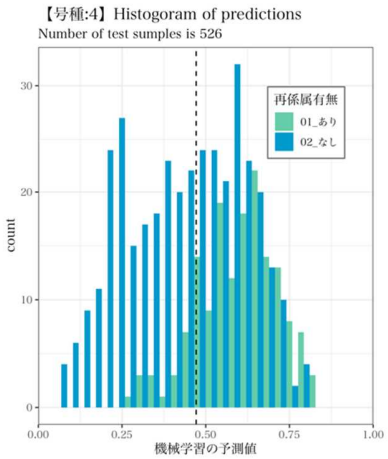

【前期区分：2008年5月から2016年5月】

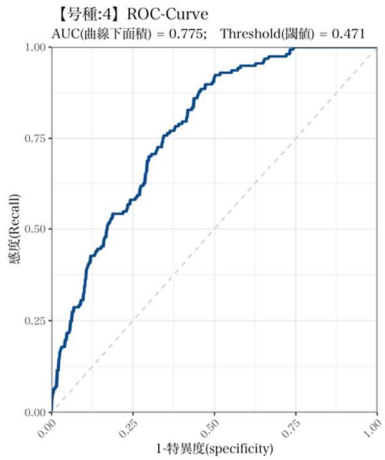



【後期区分：2016年6月（一部執行猶予制度の開始）から2017年5月】



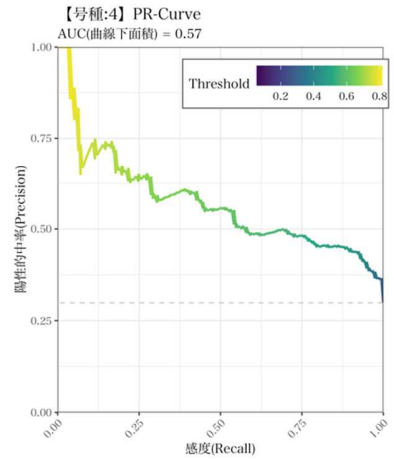



図 2.7 四号観察に該当する学習データの全事案(全期間)で学習を行った機械学習モデルに対する、前期データ(検証用データの一部)、後期データ(検証用データの一部)、全期間データ(検証用データの全事案)における予測結果。異なる列は左から順に、機械学習の予測値分布、ROC 曲線、PR 曲線を示す。異なる行は上から順に、前期データ、後期データ、全期間データに対する結果を示す。

## 2.4 結果の解釈

本章の解析の結果、基礎情報を用いた 5 年以内再係属の予測性能と項目ごとの予測的貢献度が得られた。また、三号観察と四号観察の対象データについては、それぞれの号種ごとに、前期区分(2008 年 5 月から 2016 年 5 月まで)と後期区分(2016 年 6 月から 2017 年 5 月まで)の対象データに分けて解析を行い、それぞれの予測性能を得た。

各号種にデータを分けて機械学習モデルを構築したところ、一号観察から四号観察、全号種の全てのデータ区分において一定の予測性能が得られた。具体的には、AUC-ROC が全号種：0.747、一号観察：0.738、二号観察：0.717、三号観察：0.760、四号観察：0.713 となった。これらの数値は、単純な数値比較は困難であるものの、保険数理的アプローチにより CFP データから再係属を予測した先行研究(羽間・勝田, 2021)が示した性能指標値(AUC-ROC = 0.650)よりも高い。現時点では基礎情報とリスク予測モデリングを用いることで、既存の手法より、高い精度での再係属予測が可能であることが示唆された。試験的にモデルを導入する際は、本結果を見る限り号種別にモデルを構築しても、レコード数や精度の面では不足がないように思われる。なお、2.3.3 章でも述べたように、本章の解析で用いた各種予測性能指標はアウトカム(本章では再係属)の該当率によって、解釈に注意を要する部分があることに留意されたい。

各号種区分における項目の予測的貢献度は、どの号種においても、「X18_罪名.非行名 1」が比較的上位に挙がることが観測された。実際にリスク予測モデルを現場へ導入する際は、SHAP 等による貢献度を参考に項目の選定が可能である。許容される精度が保たれる最小の項目数でモデルを構築することで、現場の職員が入力すべき項目数を減らし、業務の負担軽減、効率化に貢献することが期待される。なお、SHAP はあくまで予測的貢献度を測る指標であって、項目とアウトカムの因果関係や介入すべき項目に関しての示唆を与える指標ではないことに留意されたい。

三号観察と四号観察の対象データについては前期区分と後期区分にデータを分けて解析を実施した。全期間のデータで学習を行った後、前期区分データ、後期区分データ、全期間データのそれぞれで精度評価を実施したところ、三号観察と四号観察共に、後期区分データにおける AUC-PR の値が全期間データ、前期区分データのそれを上回っていることが観測された。ここから、後期区分データに対しては、再係属の予測がより容易である可能性が示唆された。ただし、データ区分間でのサンプル数の差異、アウトカムの定義の差異などが要因になっている可能性も十分にあることに留意されたい。また、検証用データが異なるため、予測性能指標の単純な比較が正当化されないことにも注意されたい。

次に、後期区分の学習データを使用して機械学習モデルを構築し、後期区分の検証用データで精度評価を行った。その結果、AUC-PR や AUC-ROC などの総合的な精度評価指標は、全期間データで学習を行った際の結果に比べ低下した。学習データのレコード数の低下が一つの大きな要因であると考えら

れる。ここから、実際の現場にリスク予測モデルを導入する際は、データの経年変化を考慮した期間区分に加え、データのサンプル数なども鑑みて、様々なデータ区分でモデルを構築し性能評価することが必要であると思われる。また、1.4.1 章で述べた「予測的妥当性」、「代表性」、「循環性・更新可能性」などの観点とも関連する事項として、リスク予測モデルを導入した後に AI モデルを更新する際は、学習データの区分と予測対象のデータの関係性を十分に吟味して更新を実施する必要がある。

# 第三章　CFP データの基礎検討と短期的再係属との関連

本章では、アセスメントツールである CFP を用いて蓄積された、各種データの概況を整理するとともに、CFP データが再係属の予測にどの程度貢献するか、貢献する情報にはどのようなものがあるかについて、基礎検証を実施し、その結果を報告する。なお、CFP データには、統計的分析の結果を意味する順序尺度に相当する変数や、対象者の属性に相当する項目への該当情報、対象者および対象者周辺の特徴領域ごとに整理されたテキスト情報(以下、要因テキスト)などが含まれる。ここでは、特にテキスト情報に内包された特徴の整理と、CFP データ全体を用いた予測への利活用可能性について焦点を当てた検討を行う。

## 3.1 解析の目的と検証課題

本章でのデータ解析の目的は、次の二点である。

(1) 蓄積された CFP データ、特に要因テキストの概況を把握する

(2) CFP データが再係属の予測に貢献する度合いを把握する

当該目的を達成するために、本章で実施する検証課題を次の通り設定した(表 3.1, 表 3.2)。

表 3.1 CFP データ解析の目的と検証課題 1

| 解析目的(1)：蓄積された CFP データ、特に要因テキストの概況を把握する |
|---|
| 【具体的検証課題】<br>・CFP データに含まれる要因テキストについて、出現する単語を抽出・集計し、要因ごとに含まれるテキスト情報の概況を可視化する。<br><br>・CFP データに含まれる要因テキストに一定頻度以上で出現する単語を抽出し、「強み」と「問題」の文脈情報(CFP アセスメントツール内で用いられているもの)を付与した上で、それらと再係属との関連性を、L1 正則化回帰モデルを用いて簡易的に評価する。 |

表 3.2 CFP データ解析の目的と検証課題 2

| 解析目的(2)：CFP データが再係属の予測に貢献する度合いを把握する |
|---|
| 【具体的検証課題】<br>・CFP を用いたアセスメントによるデータ蓄積が開始された時期以降において定義可能な再係属に対 |

して、(a)事案の基礎情報のみ、(b)CFP データのみ、(c)基礎情報と CFP データの両方を用いた場合の機械学習予測を実施し、それぞれの性能を比較することで、CFP データが再係属予測に貢献する程度を評価する。

・項目貢献度指標の一種である Shapley Additive Explanations を用いて、機械学習を用いた解析に耐えうる例数および再係属該当率を有する保護観察の号種区分において、再係属予測に貢献する CFP 情報を評価する。

## 3.2 方法(データ抽出)

### 3.2.1 データ源からのデータ抽出

本章で利用したデータは大別して 2 種類ある。

第一に、事件管理システムデータベースより、事業依頼元担当者が、CFP によるアセスメントデータが格納された 4 つのテーブルについて、その全件を抽出した。対象は、「CFP 開始時評定情報」、「CFP 開始時明細情報」、「CFP 要因情報」、「CFP 要因分析情報」を格納した 4 つのテーブルである。当該テーブルの抽出は、2022 年 5 月 10 日に実施された。

第二に、保護観察対象者に関連するデータを結合したテーブルである「調査研究用データ」は、事件管理システムに含まれる複数テーブルの情報を元に、事業依頼元管理のもと、事業依頼元の担当者によって、2022 年 5 月 10 日に抽出(結合処理を含む)された。これらのテーブルの基礎情報は次の通りである。

表 3.3 データ源からのレコード抽出

| 対象テーブル | レコードサイズ | 項目数 | テーブル概要 |
|---|---|---|---|
| CFP 開始時評定 | 24,198 | 13 | ID や日時等の管理項目に加えて、CFP 統計分析の結果に相当する評定区分(低い・中程度・高い)が格納されたテーブル |
| CFP 開始時明細 | 103,106 | 12 | ID や日時等の管理項目に加えて、対象者の属性に関する 16 のラベルについての該当情報が格納されたテーブル |
| CFP 要因 | 320,343 | 13 | ID や日時等の管理項目に加えて、CFP の 8 つの要因ごとに、対応するアセスメント情報(テキスト)が格納されたテーブル |
| CFP 要因分析 | 258,629 | 11 | ID や日時等の管理項目に加えて、要因ごとに記載された内容が「強み」あるいは「問題」に相当するかの文脈情報が格納されたテーブル |

| | | | |
|---|---|---|---|
| 【再掲】調査研究用データ | 432,721 | 462 | 本事業調査研究用に事前結合された、対象者の事案に関連する各種基礎情報が格納されたテーブル |

本作業では、データ源からあらかじめ抽出された表 3.3 のテーブルを対象に、以降のデータ前処理や解析作業等の手続きを実施した。

### 3.2.2 要因テキスト概況把握対象レコードの抽出と結合

本章の解析目的(1)に設定した要因テキストの基礎集計に組み入れる際の適格基準には、「保護観察期間開始日が 2021 年 1 月 1 日以降であること」、「削除フラグの項目に該当ラベルが付与されているレコードは除外すること」の 2 つを設定した。

CFP についてのデータが格納された 4 つのテーブルは、次の手続きを経た上で、各テーブルに格納された裁判所身柄 ID と保護観察期間開始日をキーとして結合した。(a)CFP 開始時評定テーブルは結合キーを除いて評定区分のみを抽出し、(b)CFP 開始時明細テーブルはロング型で格納された 16 のラベル情報について裁判所身柄 ID および保護観察期間開始日の組み合わせをユニークとしたダミー変数に展開し、(c)CFP 要因テーブルについては 3.3.1 節に記載する形態素解析を実施した後に、当該処理によって作成した単語頻度等の変数を裁判身柄 ID と保護観察期間開始日、さらに要因区分をユニークとするレコードに対して結合し、(d)CFP 要因分析テーブルは、(c)に対して裁判身柄 ID と保護観察期間開始日、さらに要因区分をキーに結合を実施した後、要因区分ごとに得られた単語頻度情報に対して文脈情報(強み・問題)を付与した。(e)そして最後に、裁判所身柄 ID と保護観察期間開始日の組み合わせでユニークとなるよう、要因区分・文脈情報・出現単語頻度の情報が合成表現された項目で展開したデータを作成した後に、CFP 関連テーブルを全て結合した。これにより、裁判所身柄 ID と保護観察期間開始日をユニークとし、要因区分・文脈情報・出現単語別の項目が列となり、フィールドに出現度数をとる「CFP テーブル」を作成した。

### 3.2.3 再係属予測におけるアウトカム定義・区分設定・レコードと項目抽出

再係属の定義は、前節に示した「CFP テーブル」の裁判所身柄 ID と保護観察期間開始日で「調査研究用データ」との結合を実施したのちに、第二章の解析と同様である「保護観察開始日から 5 年以内に再度、犯罪又は非行をして保護観察所に係属すること」との定義で再犯・再非行ラベルが付与されたテーブル情報を結合することで定義した(本来的には日付情報等を用いたデータ上での厳密な定義を行う必要があるが、保護観察の号種移動など、実質的な再犯・再非行を受託者が受領したデータセットから定義することが困難であったため、別途受領していた 5 年以内の再犯・再非行のラベルを使用する選択となった)。その上で、183 日間(約半年以内)のフォローアップを設けることで、「半年から 1 年 5 か月以内の再係属」を定義した(以下、短期的再係属)。短期的再係属の予測に際して、項目貢献度等の指標は、少なくとも「保護観察の号種」によって解釈のあり方が異なる可能性がある。よって、本解析では、号種 1(いわゆる一号観察)から号種 4 までのそれぞれについて解析を実施することを前提に、号種でのサブグループ設定を行った。

短期的再係属の予測には、(a) データ結合によって生じた同一対象者 ID における同一保護観察期間開始日のレコードを除外し、(b)短期的再係属ラベルを付与したのちに、再係属の半年間の追跡期間を

設けるため、2021 年 10 月 1 日以降に保護観察期間開始日が付与されたレコードを除外し、(c)対象外の号種(五号観察)が含まれるレコードを除外したデータを組み入れた。基礎情報に関する項目については、第二章の方針に準じて、(1)全てのレコードが欠損している項目または 99%以上を目安として著しい欠損が含まれる項目を除外し、(2)保護観察開始前の各種情報、保護観察開始時、保護観察期間中に取得可能な項目を組み入れ対象とし、(3)予測対象と明らかに無関係であると解析者によって判断された日時や ID 等の項目を除外する、という条件を設定した。また、他の特徴量として、再犯・再非行までの定義期間条件が異なる対象者情報が含まれたことから、その条件不一致に対する調整項目として、抽出時期始点である 2021 年 1 月 1 日を「1」とし、経過日数での連番を付与した特徴量を追加した(date_seq 変数)。また、CFP 要因分析テーブルから、対象者ごとに「どの要因に強み・問題のラベルが付与されていたか(Strength_…,Problem…)」を指示する変数と、対象者ごとの強み・問題の記入個数(CountStrength, CountWeakness 変数)、強みと問題の記入個数の差と合計(CountDiff, CountAll)の変数を作成し、予測に使用する特徴量として解析対象に含めた。

## 3.3　方法(データ解析)

### 3.3.1 形態素解析

形態素解析とは、テキストを品詞の単位に分解する手法である。本章解析では、CFP データに含まれる８つの要因(０１：家庭、０２：家庭以外の対人関係、０３：就労・就学、０４：物質使用、０５：余暇、０６：経済状態、０７：犯罪・非行や保護観察の状況、０８：心理・精神状態)ごとに記録されたテキストについて、CFP の実施単位で形態素解析を行い、品詞分解を行った。そして、名詞(一般名詞、固有名詞、形容動詞語幹、副詞可能)と形容詞(自立)を抽出し、CFP の実施単位で出現頻度を計上した。形態素解析には、mecab(0.996)と辞書 mecab-ipadic(2.7.0-20070801)を使用し、実行には統計解析環境 R のパッケージ RMeCab(version 1.10, Ishida & Kudo, 2022)を使用した。なお、テキスト情報中に含まれる単語の品詞分類の結果は、使用する辞書等のバージョンなど、解析の実行環境に依存する点に留意されたい。

### 3.3.2 L1 正則化回帰モデル

L1 正則化回帰モデルとは、回帰モデルのパラメータに L1 ノルムの罰則項を付与することにより、予測に使用する項目選択と、過学習に配慮した予測を同時に実現する解析手法である。罰則の大きさによって選抜される項目の数と内容が変化し、得られる予測性能も変化する。回帰モデルは、「項目の情報を重み付け、その合計値を目的変数の予測に利用する」という基本構造を持つ。妥当性はモデル自体が持つ仮定や得られた予測性能に依存するものの、項目に対する重み(係数)の値によって、各変数と予測対象との関係を捉えることができる。本章では、CFP 要因テキストに含まれた単語頻度情報から、再係属と関連しうるものを抽出する際に使用した。交差検証法により誤差最小となった罰則項の大きさを採用し、当該条件で選抜された変数の係数を抽出した。CFP 要因テキストの概況を多面的に把握する際の手段の一つとして利用し、選抜された項目や係数の大きさについては、全て仮説を得る前段階としての暫定的な解釈にとどめた。また、選抜項目について解釈する上では、保護観察の区分ごとに層別化したデータで解析を実施する必要があると考えられることから、当該解析は号種別のサブ

グループに対して実施することをあらかじめ設定した。実行環境は統計解析環境 R(version 4.2.0, R Core Team, 2022)を使用し、L1 正則化回帰モデル(再係属の有無が二値変数であるためロジスティックモデル)の実行は glmnet パッケージ(version 4.1.4, Friedman et al., 2011)を使用した。

### 3.3.3 LightGBM

再係属の有無を予測するために、機械学習モデルの一つである LightGBM(第二章参照)を使用した(Guolin Ke el al., 2017)。学習(最適化)の目的関数には対数損失(logloss)を採用し、クラス不均衡に対するモデル内対処を有効とした。具体的には、neg_bagging_fraction を学習データにおける 再係属率/(1-再係属率)とした。また評価関数(metrics)には ROC 曲線下面積を使用した。チューニング対象としたハイパーパラメータは、max_depth={-1, 2, 3, 6, 8}、num_leaves = {20, 30, 40}、lambda_l1 = {24, 28, 30, 32}、learning_rate = {0.01, 0.05, 0.1}、min_data_in_leaf = {10, 20, 30, 40}であり、それぞれの組み合わせを使用し、グリッドサーチで評価関数が最小となるものを求めた。このとき、num_iteration = 600, early stopping rounds = 50 とし、交差検証(3-fold)によって最適な round 数を求め、当該 round 数での最終的な学習を行った。他のハイパーパラメータは全てパッケージのデフォルト値を使用した。解析の実行には、統計解析環境 R(version 4.2.0, R Core Team, 2022)と、lightgbm パッケージ(version 3.3.2, Shi et al., 2022)を使用した。

LightGBM の適用は、第二章に準じて号種別での実施を前提とし、アウトカムの著しいクラス不均衡や、例数の著しい不足等がある場合には、当該区分を適用外とする方針とした。性能評価指標についても、第二章と同様に、AUC-PR、AUC-ROC、Accuracy、感度、特異度、陽性的中率、陰性的中率を採用し、多面的に評価した。

また、本章の目的である「CFP データが再係属の予測に貢献する度合いを把握する」ため、LightGBM は(1)CFP データのみを項目として用いた場合、(2)基礎情報のみをデータとして用いた場合、(3)基礎情報と CFP データの両方を項目として用いた場合の 3 通りで実施した。

### 3.3.4 Shapley Additive Explanations

L1 正則化回帰モデルとは別に、LightGBM を用いた予測において貢献した CFP 要因テキスト情報を把握するために、SHAP(Scott and Lee, 2017)を採用した。第二章と同様に、学習データのレコードごと、かつ、予測に使用した項目ごとに得られた SHAP について、その絶対値を項目ごとに平均した値を使用した。この時、各項目に対する SHAP の値を解釈する上では、保護観察の区分ごとに層別化した解析を実施する必要があると考えられる。したがって、号種別での LightGBM の適用結果について指標の出力を行うこととした。

## 3.4 結果(CFP データの基礎評価)

### 3.4.1 データ抽出・結合の結果

適格基準に基づく解析データ抽出の結果、「CFP 要因分析」テーブルについては、削除フラグの項目を有していたため、当該列に該当ラベルが付与された 67 レコードが除外された。その後、データ結合処理を行った結果、形態素解析の結果抽出された文脈つき単語 8588 種が得られ、当該単語のいずれか

に該当のあったユニーク ID 数は 9441 となった。ここで、8588 種の文脈つき単語を展開した場合、該当フィールド数が著しく低い項目を含む、疎なデータ行列が得られることとなる。本作業では、データの概況把握を行う初期評価であるという前提・目的に照らして、レコードサイズに対して過大な項目数を取得することを避け、9441 件のレコードにおいて 1000 回以上出現した単語のみを展開した。単語の種類別の出現頻度情報を図 3.1 に示す。

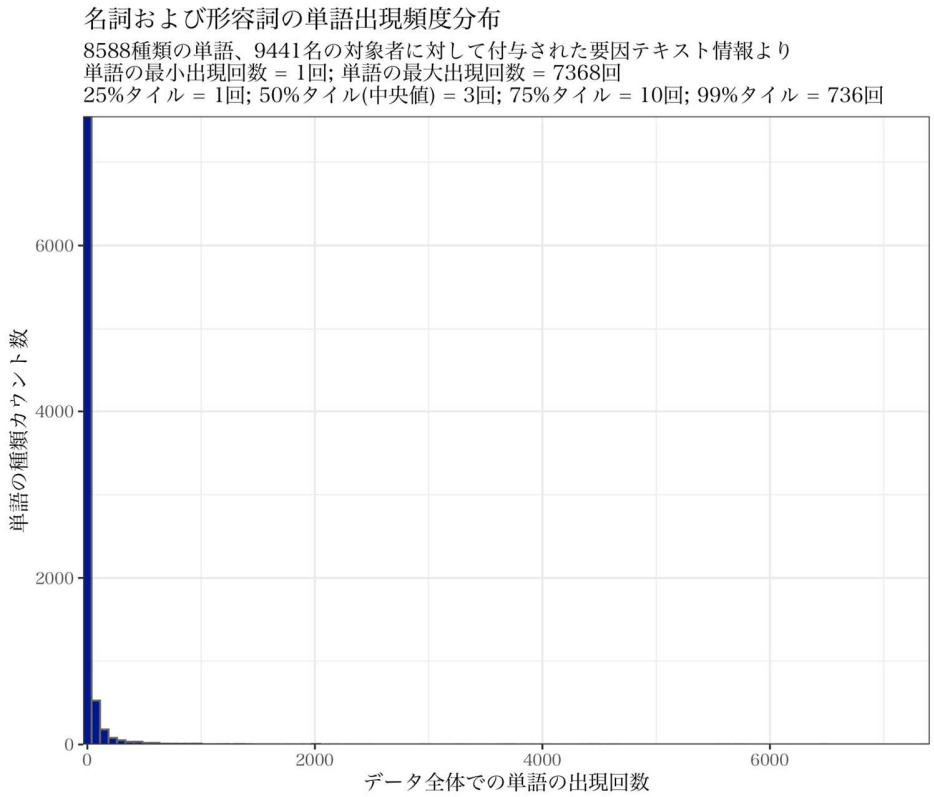



図 3.1 CFP 要因テキストデータの出現頻度度数分布と要約情報

名詞(一般・固有・形容動詞語幹・副詞可能)と形容詞(自立)のみを抽出したもの

(出現回数が少ない単語の数が多く、頻繁に使用される単語が少ない)

その結果、31 の文脈つき単語と、対象用語が一つ以上含まれたユニーク ID 数は 8605 件となった。当該データを、事案の基礎データに相当する調査研究用テーブルと結合し、同一対象者において保護観察期間開始日の重複するレコードの後者を抽出したところ、24198 件の結合データが得られた。追跡期間の保護観察期間開始日が付与されたレコードを除外した結果、解析対象レコードは 13697 件となった。当該データにおける短期再係属の該当率は 5.3%(727 件/12970 件)となり、号種別の係属率は号種 1 で 8.25%(3745 件中)、号種 2 で 8.48%(1120 件中)、号種 3 で 3.41%(7211 件中)、号種 4 で 4.75%(1621 件中)となった。当該区分別該当率から、全ての号種について、号種別での解析を実施することとした。短期再係属予測の解析に組み入れた項目数は、目的変数と号種情報を除いて、基礎データと CFP 情報を含めて 120 項目となった。概要を表 3.4 に示す。

表 3.4 短期再係属予測に組み入れられた項目情報(目的変数と号種情報の 2 項目を除く)

| 領域 | 区分等 | 項目数 | 概要 |
|---|---|---|---|
| 基礎データ | 対象者基礎情報 | 12 | 保護観察期間開始時の年齢や性別、生活歴等に関する基礎情報 |
| | 処遇等基礎情報 | 40 | 事件や罪名、刑や入所歴、保護観察の処遇、処遇等の実施機関に関する基礎情報 |
| CFP データ | 評定区分 | 1 | 開始時統計的評価情報 |
| | CFP チェック項目 | 16 | 主要なアセスメント観点に関する該当情報 |
| | CFP 要因テキスト情報 | 31 | 対象データにおいて、CFP 要因テキスト内で 1000 回以上出現した単語(名詞・形容詞)に「強み」と「問題」の文脈情報を付与したユニークな単語の組に対する該当情報(補足資料表 S3.1 に内容記載) |
| | CFP 要因分析情報 | 16 | 8 つの要因区分に対して、「強み」と「問題」のラベルが付与されていたか否かに関する情報 |
| | 追加生成項目 | 5 | ・対象者ごとに算出された「強み」と「問題」それぞれのラベル該当数、ラベル該当数の和と差に関する情報。<br>・対象者で再犯・再非行の定義期間条件が異なることに対する調整項目(日付連番) |

### 3.4.2 要因別テキストにおける出現単語の集計

CFP 要因に含まれた 8 つの区分における、単語情報について集計した。その結果、要因区分別で表 3.5 に要約されうる単語が出現回数の上位に含まれた(上位 100 の単語詳細は、補足資料 Figure S3.1 を参照)。

表 3.5 CFP 要因テキストに出現した上位単語の概要(解析者による要約整理)

※名詞(一般、固有、形容動詞語幹、副詞可能)および形容詞(自立)のみを抽出して計上

| CFP 要因区分 | 上位出現単語の内容(解析者による任意の要約) |
|---|---|
| ０１：家庭 | 家族成員の続柄が高頻度で出現した他、家族関係の様相を形容する単語、生活状況や障害、暴力等に関連する単語が含まれた。犯罪の種別に相当する単語なども一部含まれた。 |
| ０２：家庭以外の対人関係 | 対象者の生活歴の中で交友等が発生したと思われる様々な関係者(グループ)の属性や、対象者との関係性を形容する単語が上位に含まれた。特定の犯罪種別にまつわる用語や、インターネット等を介した関係性を示唆する単語も含まれた。 |
| ０３：就労・就学 | 就学先や就労先に相当する単語や、職種や雇用形態に関する用語、時期や期間に関する単語、就学・就労先の関係者の属性を示す用語、対象者の意欲や態度を形容するものと思われる単語が含まれた。 |
| ０４：物質使用 | 乱用されたと思われる物質の種別や、物質使用動機や使用の誘発に関わると思われる条件、物質使用環境や頻度、関係者や治療などに関連する単語が含まれた。 |
| ０５：余暇 | 余暇の有無自体を示したと思われる用語、余暇の内容、余暇の時間を過ごす関係者の属性に関する用語が抽出された。ギャンブルや接待飲食、性風俗関連の用語、特定の犯罪種別にまつわる用語が一定数含まれた。 |
| ０６：経済状態 | 経済水準を形容する単語のほか、衣食住に関連する用語、生計単位に含まれうる関係者の続柄、経済的負債や金銭管理・使い方に関連すると思われる用語、特定の犯罪種別にまつわる用語などが含まれた。2020 年 4 月ごろより流行した新型コロナウイルス感染症に関する用語なども含まれた。 |
| ０７：犯罪・非行や保護観察の状況 | 罪名や刑罰・処遇に関する用語、対象者の態度を形容する用語のほか、対象者の関係者や、保護観察中に発生したと思われる様々な種類の出来事に関連する用語が含まれた。 |
| ０８：心理・精神状態 | 知的側面、情緒的側面、心理症状や障害に関する用語のほか、意志や意欲、社会的側面、その他ポジティブ・ネガティブに形容する用語、特定の犯罪種別にまつわる用語などが含まれた。 |

### 3.4.3 抽出した利用単語と再係属との関連について(L1 正則化回帰モデル)

CFP データのみを用いた号種別のデータ区分に対する L1 正則化回帰(ロジスティック)の結果、表 3.6 に示す通りの予測性能(AUC-PR)となった。検証データに対する予測性能が十分に得られていないことから、学習データで求められた選抜係数とその係数値に解釈上の意義は小さく、誤解を招く恐れから、掲載を割愛する。

表 3.6 号種別データに適用した L1 正則化回帰モデルの予測性能(AUC-PR)

| 対象データ | PR 曲線下面積<br>(AUC-PR) | 基準値(検証データでの短期再係属率) | 選抜項目数<br>(切片を含む) |
|---|---|---|---|
| 号種 1 | 0.13 | 0.10(10%) | 11 |
| 号種 2 | 0.14 | 0.12(12%) | 3 |
| 号種 3 | 0.07 | 0.05(5%) | 7 |
| 号種 4 | 0.05 | 0.05(5%) | 10 |
| 全号種 | 0.09 | 0.06(6%) | 33 |

## 3.5 結果(機械学習)

LightGBM による短期再係属予測の結果、号種別および全号種で各種の性能指標が得られた。1 号観察から 4 号観察、そして全号種の結果を表 3.7〜表 3.11 に示す。なお、性能指標に関する具体的な解釈例は、補足資料 S2 の記載例を確認されたい。

表 3.7 号種 1 における異なる項目条件での予測性能指標一覧

AUC-ROC には 95%信頼区間、AUC-PR には基準値(検証データの短期再係属該当率)を付記。閾値は感度と特異度の和が最大となるものを記載しており、Accuracy から Negative Predictive Value までの指標は、当該閾値を用いて算出している

| 性能指標 | 基礎データのみ | CFP データのみ | 基礎データ+CFP |
|---|---|---|---|
| AUC-ROC | 0.706[0.644, 0.767] | 0.663[0.600, 0.726] | 0.718[0.658, 0.777] |
| AUC-PR | 0.243(0.100) | 0.169(0.100) | 0.239(0.100) |
| Threshold(閾値) | 0.530 | 0.354 | 0.466 |
| Accuracy | 0.732 | 0.593 | 0.641 |
| Recall(感度) | 0.569 | 0.694 | 0.722 |
| Specificity(特異度) | 0.749 | 0.582 | 0.632 |

| | | | |
|---|---|---|---|
| Precision | 0.194 | 0.150 | 0.173 |
| Negative Pred.value | 0.942 | 0.947 | 0.955 |

表 3.8 号種 2 における異なる項目条件での予測性能指標一覧

AUC-ROC には 95%信頼区間、AUC-PR には基準値(検証データの短期再係属該当率)を付記。閾値は感度と特異度の和が最大となるものを記載しており、Accuracy から Negative Predictive Value までの指標は、当該閾値を用いて算出している

| 性能指標 | 基礎データのみ | CFP データのみ | 基礎データ+CFP |
|---|---|---|---|
| AUC-ROC | 0.733[0.649, 0.817] | 0.668[0.593, 0.743] | 0.740[0.66, 0.819] |
| AUC-PR | 0.218(0.120) | 0.171(0.120) | 0.200(0.120) |
| Threshold(閾値) | 0.415 | 0.167 | 0.341 |
| Accuracy | 0.603 | 0.504 | 0.576 |
| Recall(感度) | 0.885 | 0.885 | 0.923 |
| Specificity(特異度) | 0.566 | 0.455 | 0.530 |
| Precision | 0.211 | 0.176 | 0.205 |
| Negative Pred.value | 0.974 | 0.968 | 0.981 |

表 3.9 号種 3 における異なる項目条件での予測性能指標一覧

AUC-ROC には 95%信頼区間、AUC-PR には基準値(検証データの短期再係属該当率)を付記。閾値は感度と特異度の和が最大となるものを記載しており、Accuracy から Negative Predictive Value までの指標は、当該閾値を用いて算出している

| 性能指標 | 基礎データのみ | CFP データのみ | 基礎データ+CFP |
|---|---|---|---|
| AUC-ROC | 0.762[0.705, 0.819] | 0.657[0.592, 0.722] | 0.764[0.708, 0.819] |
| AUC-PR | 0.142(0.05) | 0.087(0.05) | 0.139(0.05) |
| Threshold(閾値) | 0.493 | 0.306 | 0.416 |
| Accuracy | 0.749 | 0.548 | 0.640 |
| Recall(感度) | 0.662 | 0.723 | 0.800 |
| Specificity(特異度) | 0.753 | 0.540 | 0.633 |
| Precision | 0.112 | 0.069 | 0.093 |

| | | | |
|---|---|---|---|
| Negative Pred.value | 0.979 | 0.976 | 0.985 |

表 3.10 号種 4 における異なる項目条件での予測性能指標一覧

AUC-ROC には 95%信頼区間、AUC-PR には基準値(検証データの短期再係属該当率)を付記。閾値は感度と特異度の和が最大となるものを記載しており、Accuracy から Negative Predictive Value までの指標は、当該閾値を用いて算出している

| 性能指標 | 基礎データのみ | CFP データのみ | 基礎データ+CFP |
|---|---|---|---|
| AUC-ROC | 0.652[0.494, 0.810] | 0.687[0.551, 0.823] | 0.692[0.542, 0.841] |
| AUC-PR | 0.092(0.05) | 0.137(0.05) | 0.095(0.05) |
| Threshold(閾値) | 0.428 | 0.451 | 0.492 |
| Accuracy | 0.762 | 0.512 | 0.858 |
| Recall(感度) | 0.533 | 0.800 | 0.533 |
| Specificity(特異度) | 0.773 | 0.498 | 0.874 |
| Precision | 0.103 | 0.072 | 0.170 |
| Negative Pred.value | 0.972 | 0.981 | 0.975 |

表 3.11 全号種における異なる項目条件での予測性能指標一覧

AUC-ROC には 95%信頼区間、AUC-PR には基準値(検証データの短期再係属該当率)を付記。閾値は感度と特異度の和が最大となるものを記載しており、Accuracy から Negative Predictive Value までの指標は、当該閾値を用いて算出している

| 性能指標 | 基礎データのみ | CFP データのみ | 基礎データ+CFP |
|---|---|---|---|
| AUC-ROC | 0.731[0.692, 0.770] | 0.658[0.614, 0.702] | 0.726[0.687, 0.764] |
| AUC-PR | 0.146(0.06) | 0.115(0.06) | 0.160(0.06) |
| Threshold(閾値) | 0.447 | 0.509 | 0.349 |
| Accuracy | 0.654 | 0.624 | 0.573 |
| Recall(感度) | 0.706 | 0.595 | 0.765 |
| Specificity(特異度) | 0.651 | 0.626 | 0.561 |

| | | | |
|---|---|---|---|
| Precision | 0.107 | 0.086 | 0.094 |
| Negative Pred.value | 0.974 | 0.963 | 0.976 |

表 3.7〜表 3.11 に示した各解析結果で得られた、項目貢献度指標(SHAP)に関する情報は、補足資料 S3.2 ~ S3.16 を参照されたい。

## 3.6 結果の解釈

本章の解析の結果、CFP 要因テキストデータに出現する単語の基礎情報と、現時点での CFP データによる短期再係属の予測性能が得られた。

各単語がデータ内で出現した回数は、名詞と形容詞の一部に絞った場合、単語の種類によって大きな開きがあることが示された。家族等の続柄を示す単語や、犯罪の種別を示す単語などの出現頻度が多い一方で、その他多くの単語は使用頻度が著しく低かった。

CFP データを用いた短期再係属予測の結果は、CFP データ単体であっても、一定の予測性能が得られた。具体的には、CFP データのみの場合で「0.657(号種 3) ≦ AUC-ROCs ≦ 0.687(号種 4)」の値が得られ、CFP と基礎データを組み合わせた場合には「0.692(号種 4) ≦ AUC-ROCs ≦ 0.764(号種 3)」の性能指標値が得られた。これらは、データが異なるために単純な数値比較は困難であるものの、保険数理的アプローチを用いた類似する先行研究(羽間・勝田, 2021)が示した性能指標値(AUC-ROC = 0.650)よりも高い。従来と異なる解析的アプローチ(リスク予測モデリング)を用いることや、CFP を用いて得られる情報を効果的な形式で蓄積することで、より高い精度での再犯・再非行の予測が実現できる可能性がある。

なお、感度および特異度を中心的な指標としてモデルや検査を評価する場合、予測対象となるアウトカムの該当率が低ければ、陽性的中率(Precision, Positive Predictive Value)が「印象よりも」低い値となることが知られている(ベースレートの誤謬)。リスク予測モデルを実装する際には、陽性的中率の低下による偽陽性の増大に十分な留意が必要となる。

CFP データの今後の発展については、第一に、テキスト情報等からさまざまなアセスメント観点を抽出し、それを項目化するという手段が考えられる。予測的妥当性に貢献する項目や、介入によって変容可能な項目(介入標的になる反応性をもつ項目)を設置することで、テキストの記録者等によって情報の表現形式が変化しない信頼性のあるデータを蓄積することにつながる。アセスメント観点を項目化することにより、いくつかの発展的な利用にもつながると考えられる。

一例として、CFP に基づくアセスメントは保護観察期間の開始時だけでなく、経過中にも実施可能性がある。介入標的となるアセスメント項目を設けておくことで、各種処遇を経て変化する対象者の状態の経時的な変化も追うことが可能となる。そして、変化の度合いをも情報として活用することで、将来的な再犯・再非行の予測的評価がより高い精度で得られる可能性がある。他の例として、介入標的となるアセスメント項目を設けておくことで、「対象者の特定の強みを伸ばす支援や介入、問題を改善する支援・介入によって、どの程度の再犯・再非行の低減効果が見込めるか」といった解析技術の利用や、結果の活用方法への発展が期待できる。こういった、「事前に介入効果が大きいと推測される介入標的に対して処遇プログラムを展開してゆく」といったアプローチを実現する方法がすでに

いくつか提案されてきている。技術の概要は、第四章の 4.2.1 節 「Counterfactual explanations」に記載する。また、多面的なアセスメント項目を抽出し、その予測的妥当性を評価する際の一つの手段として想定される調査技術は、第四章の 4.2.2 節「計画的欠損データデザイン」に示す。

# 第四章　データ解析から得られた知見と課題の整理

本章では、第二章及び第三章の解析で得られた結果を踏まえ、特に蓄積データ設計とリスク予測モデル構築に関する点について、後続した検討が必要と思われる課題点や将来的展開と、それに対応しうる方法論についての一例を整理する。

## 4.1 データ解析から得られた知見

### 4.1.1 事件管理システムのデータを用いた 5 年以内再犯・再非行予測の結果

第二章に示した解析の結果、事件管理システムに蓄積された保護観察期間開始時に取得可能な基礎情報のみを活用した場合であっても、一定の精度で 5 年以内の再犯・再非行の予測が可能であることが示された。保護観察の号種別でも、区分を考慮せず全体のデータを用いた場合であっても、保険数理的アプローチに基づく暫定の予測精度(羽間・勝田, 2021)より高い値が得られている。「既存の方法論よりも性能が良い」というある種の予測的妥当性が示されたことは、リスク予測モデルの導入根拠を補強する資料となるだろう。また、犯罪や非行、あるいはそれらに対する処遇のあり方についての経年的な質の変化を考慮した性能評価の手続きや、保護観察の号種別という条件での試験的なモデル構築を実施するという過程において、学習用データのレコードサイズや項目数等に目立った不足は観察されなかった。リスク予測モデルの構築に必要な学習データの基礎条件は、現時点でも満たしているものと考えられる。

ただし、未だ精度向上の余地があるという意味において、本解析で満足な性能が得られたとは評価されない。また、表 1.1 および表 1.2 に示した各種の検討事項を参照すれば、予測的妥当性が担保されただけでは十分であるとは言えない。後続の検証課題と発展の可能性を想定し、継続したモデル設計と評価の手続きが必要であると考えられる。

### 4.1.2 CFP データの基礎検討と短期的再係属との関連

第三章に整理した CFP 要因テキストの基礎的な特徴把握の結果、CFP アセスメント情報には、(検証範囲では短期的な)再係属を予測する観点から有用な情報が含まれていることが示された。短期の再係属を対象としている点で類似する、暫定の予測精度(羽間・勝田, 2021)よりも、CFP データのみを用いた場合に得られた精度はわずかに高く、基礎データと組み合わせた場合に得られた精度は一定程度高いと形容できる結果が得られている。また、CFP 要因テキスト情報に含まれる代表的な品詞について、その単語の出現度数には種類によって大幅な開きがあり、対象者の具体的な特徴を表現する単語は出現頻度が少なく、出現頻度が多い単語はアセスメント情報としての具体性が低いことが示された。リスク予測モデリングによって予測的妥当性を担保し、他の評価観点（例えば解釈可能性・説明可能性や非差別性など）を充足させるためには、単語情報ではなく、アセスメント観点等を項目化した方

が、利便性が高く、評価観点との整合性が高い可能性が指摘された。CFP 要因テキストに記載された各種の観点や既存の研究情報等を活用し、(1)再犯・再非行の予測に有用である、(2)再犯・再非行の主要な要因に(背景理論に基づいて)なりうる、(3)各種処遇によって変化が期待できる介入可能性と反応性をもつ、(4)評価者による評定のブレが少なく信頼性がある、といった観点を参照基準として項目を選抜し、それらの該当情報等を蓄積してゆくことが望まれるものと考えられる。

## 4.2 リスク予測モデルの発展に向けた技術的工夫の例

### 4.2.1 Shapley Additive Explanations を用いた事案単位の解釈補助

再犯・再非行のリスク予測モデルの中核的な出力は、将来的な再犯・再非行の発生についての蓋然性の度合いを示す指標になると思われる(再犯・再非行の発生確率の推定値等)。しかし、機械学習モデル等によって得られる値の出力過程は一般に了解困難であり、「ブラックボックス」と形容されることも多い。「モデルがどのようにその値を出力したのか」に関する補助機能を付することは、モデルの解釈可能性を向上させ、処遇プログラム等による主要な介入標的を検討する上での一つの参照資源になると考えられる。第二章および第三章でも使用した Shapley Additive Explanations は、データ全般における各項目の予測貢献度を評価するだけでなく、「ここの事案についての予測値を得るにあたり、個別の項目がどの程度寄与したか」に関する要約情報を得る際にも活用することができる。当該技術を図式的に示すものとして、waterfall plot がある(Scott, 2018, 図 4.1)。モデルのベースレート(切片や基準値に相当)からスタートし、「各項目の SHAP 値を合計することで、最終的な予測値が得られている」というプロセスを可視化したものである。なお、waterfall plot に示されるような SHAP の値は、例えば「再犯・再非行が発生する原因と、その影響度の大きさ」を示すものではなく、「モデルが最終的な予測値を出力するまでのプロセス」を要約的に示したものに過ぎない点には十分留意されたい。

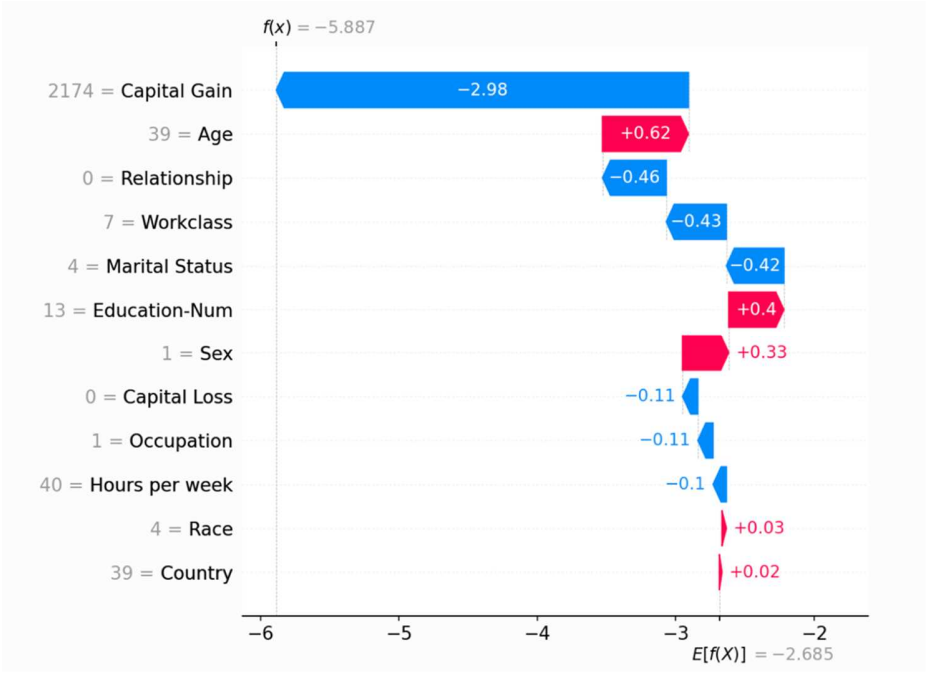


図 4.1SHAP を用いた Waterfall plot の例



引用元: Scott(2018). SHAP documentation. © Copyright 2018, Scott Lundberg. Revision 45b85c18., URL: https://shap.readthedocs.io/en/latest/example_notebooks/api_examples/plots/waterfall.html

### 4.2.2 Counterfactual Explanations

リスク予測モデルを実際の現場に導入する際は、その予測値を参照することに留まらず、事案への介入を行い、将来起こりえる負の事象を予防したいと考えるのが一般的である。予測値を変化させたい(負の事象を予防したい)際には、本事業でも用いた SHAP を参照することで、どの(介入)項目に着目すれば良いかについての検討をつけることはできる。しかし、SHAP はあくまで予測への貢献度を測る指標であり、どの項目がどの程度(介入等によって)変わった時に、どの程度予測値が変化するのかについての知見を得ることはできない。

ある事案について、再犯・再非行発生等の予測値を望まれる値にまで変化(低減等)させることを狙ったときに、元の入力(各項目の値)をどの程度変更すべきかを算出する手法として、Counterfactual Explanation(CE：反実仮想説明)と呼ばれる手法が近年盛んに研究されている(Riccardo Guidotti, 2022)。

CE の基本的な原理は、元の入力内容を徐々に変えていき、元の入力になるべく近く、かつ望まれる予測値を得ることのできる入力パターン(Counterfactual)を見つけることである。これにより、「この事例の場合、もしもこの部分がこの程度変わったら、再犯・再非行の予測値が閾値下まで低下する」という情報を得ることが可能になる。しかし、こういった「もしもこの程度まで変わったら」という反実仮想例を単純に取得しようとすると、現実には生じ得ない変更(年齢を下げるなど)が提案されてしまったり、項目間の相互作用が未考慮の Counterfactual が算出されたりしてしまう。こういった問題を緩和するために、CE 研究では様々な手法が開発・提案されてきた。

いくつか例を挙げると、Diverse Counterfactual Explanations(DICE)(Ramaravind K., *et al.*, 2020)は、複数の Counterfactual の候補を算出し、ユーザー(一般にはその分野の専門家)が実現可能と判断する Counterfactual を選択する手法である。FACE: Feasible and Actionable Counterfactual Explanations (Rafael. *et al.*, 2019)では、より現実的な項目の変化を経た(データ空間においてサンプルが密なルートを通った)Counterfactual を算出する手法である。その他にも、項目間の因果関係を考慮し、変更(介入)を施す順序も同時に提案する CE の手法なども提案されている(Kentaro K., *et al.*, 2021)。

将来的に、再犯・再非行を予測する機械学習モデルを現場へ導入する際は、上記の各種 CE が応用できると考えられる。例えば、保護観察開始時に得られた再犯・再非行の発生リスクに対して、どのような処遇を施すべきかの判断を、CE により補助できる可能性がある(ただしこの場合は実施予定の処遇や介入可能な項目をあらかじめ入力項目に含めておく必要がある)。また、保護観察期間中の急性再犯リスクを予測する機械学習モデルを導入する際は、既に実施した処遇の情報も入力に含めることで、急性再犯リスクを下げるためにどのような介入が次に必要かの提案が可能になる。

介入可能な項目の設置を検討する上での主要な観点として、(1)処遇等で変化を図る介入標的であること、(2)測定(尺度)項目に妥当性と信頼性、そして「反応性」が確保されていることを検証しておく必要がある。

(1)の介入標的であることとは、一例として、認知・行動療法等の心理的介入を処遇プログラムに用いることを想定した場合、「認知の歪み」や「行動の生起頻度」など、介入による変容を狙う対象を項目として設置することに相当する。このとき、介入標的となる対象は、再犯・再非行の発生メカニズムにおける中核的要素であることが望ましい。中核的な項目を設置することで、それが改善したときに、再犯・再非行の予測値に意味のある変化を生むことが期待できる。

(2)の妥当性と信頼性については、測定(尺度)項目自体に一般的に担保されて然るべき条件となる。なお、「反応性」とは、対象者の変化や介入等によって、測定される値が変化する性質を示す。一例となるが、完全な治癒が想定されず、症状が軽快しても常に一時的な回復状態(寛解)と見做される疾病などにおいて、「完治 / 完治していない」などの項目を設置した場合、どのような治療・介入等を施しても「完治していない」という入力しか得られないといった状況が生じる(入力に変化が生じない = 反応性がない)。他の例として、同じ測定概念であったとしても、項目の文言表現等によって、介入による得点変化が小さい尺度と大きい尺度などが存在しうる。個人の変化に応じて得点の変化が得られない(得られにくい)といった反応性の担保されない測定項目を設置した場合、必然的に、反実仮想説明への活用可能性は低下する。反応性を含めた測定(尺度)項目の評価方法は、COSMIN チェックリストが詳しい(Mokkink et al, 2019)。

### 4.2.3 計画的欠損データデザイン

リスク予測モデルの活用を前提とした時、入力される情報が評定者間で大幅に変わらないなど、入力形式が一貫し、安定していることは重要な意味を持つ。リスク予測モデルを前提として CFP データを活用することを想定するならば、CFP ツールを基礎として収集される情報にも安定性が求められる。具体的には、同じ事例であっても評定者によって記録内容が容易に異なるテキスト情報を用いるよりも、アセスメント項目を別枠で設置し、それらへの該当状況を入力する方が、入力情報の信頼性は高まると考えられる。AI-CFP の実装を目指す上で、再犯・再非行の予測に貢献するアセスメント項目の設置を検討する意義は大きい。

アセスメント項目の拡充を検討する上で、少数の原理的項目を設定するといった方針も選択肢の一つとして想定されるものの、再犯・再非行の発生を念頭に置いたアセスメントを実施する場合、考慮すべき観点は多い(多面的)と考えるのが自然だろう。その範囲は、対象者の身体・生物・心理・対社会的側面に関する個人特性や、家族関係・友人等交友関係、経済状況、生活歴など多岐にわたる。また、犯罪等の種別によって、より詳細に把握が必要な観点も異なりうる。しかし、それら全てについての測定(尺度)項目を明文化し、全ての妥当性や信頼性などを検証するには、膨大な調査コストが必要となる。先例として、児童虐待や不適切養育の発生可能性を見立てるための項目として 638 項目が一次抽出された調査報告や(髙岡他, 2022)、重篤な児童虐待行為の発生をアセスメントする項目を文献調査から 420 項目抽出したとする報告もある(髙岡他, 2020)。このような数の項目を全て組み込んだアンケート調査等を実施することは、調査協力に係る負担等を考慮しても現実的ではない。

取り扱いたい項目の規模が大きい場合に、それらを効率的に取り扱う調査手法の一つとして、計画的欠損データデザイン(Planned Missing Data Designs)がある(Kyle et al., 2020)。計画的欠損データデザインとは、統計的な欠損補完方法の活用を前提に、意図的に(完全な、あるいは、ブロック単位などでの)無作為な欠損を発生させる手法である。ある調査回答者には、大元の項目プールから無作為に抽出した項目セット A を提示し、別の調査回答者には、無作為に抽出した項目セット B の回答を依

頼するといった形式にて調査が実施される。こういった工夫により、ある一人の調査協力者が回答する項目数(縦断調査の場合には、回答するタイミングの数)を減らすことが可能となる。計画的欠損データデザインの手法にはいくつかのものが提案されており、完全にランダムな項目抽出を行うもの(random item allocation)、いくつかの項目をセットにしたブロック単位での提示を行うもの(N-block design)、時系列の反復測定調査の場合に無作為なタイミングでの測定を個々の対象者に割り当てるもの(wave design)などがある(Todd & Mijke, 2013)。児童虐待に関する先例(髙岡他 2020; 髙岡他 2022)では、random item allocation デザインを用いて、リスク予測モデルへの活用を前提に、単一の調査で数百に及ぶ全ての項目についての予測的妥当性に関する基礎評価を実施している。

計画的欠損データデザインでは、調査協力者の負担を軽減した効率的な情報収集を行うことにより、「元来生じ得た大きな調査回答負担による、不用意な回答者の脱落や欠損の発生が防止できる」という利点があるとされている(Todd & Mijke, 2013)。それだけでなく、本事業の文脈に照らせば、「最終的に採用しなかったアセスメント項目についても基礎情報が得られ、採用されたアセスメント観点の透明性や完全性(何がリスク予測モデルで考慮されており、何が考慮されていないか)などに関する議論が可能となり、項目の更新等に係る参照源としても活用できる」という複数の利点を得ることにもつながるだろう。図 4.2 および図 4.3 は、髙岡他(2022)で実施された計画的欠損データデザインの概要ならびに、ある一つの項目について、当該調査で得られた基礎情報の一例を示す。

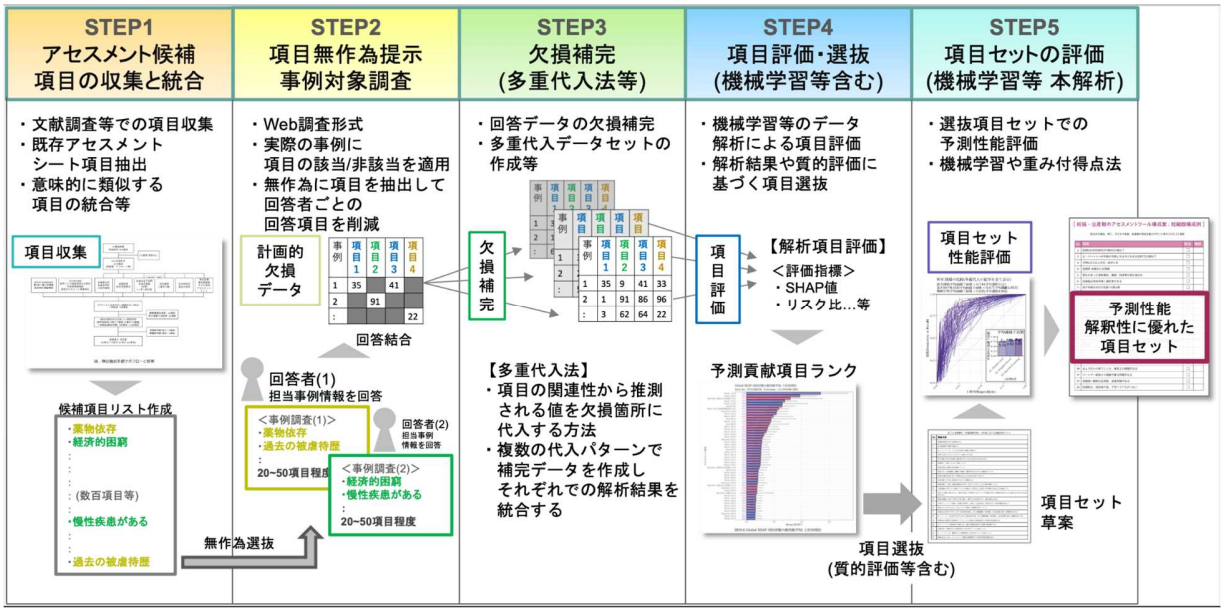



図 4.2 計画的欠損データデザインを用いた調査によるアセスメント項目抽出の例(概要図)

※ 産業技術総合研究所(2022). 令和 3 年度 子ども・子育て支援推進調査研究事業「母子保健における児童虐待予防等のためのリスクアセスメントの在り方に関する調査研究」調査事業報告書を元に受託者が作成(URL: https://staff.aist.go.jp/kota.takaoka/research/mhlw_parentChildHealth.html)

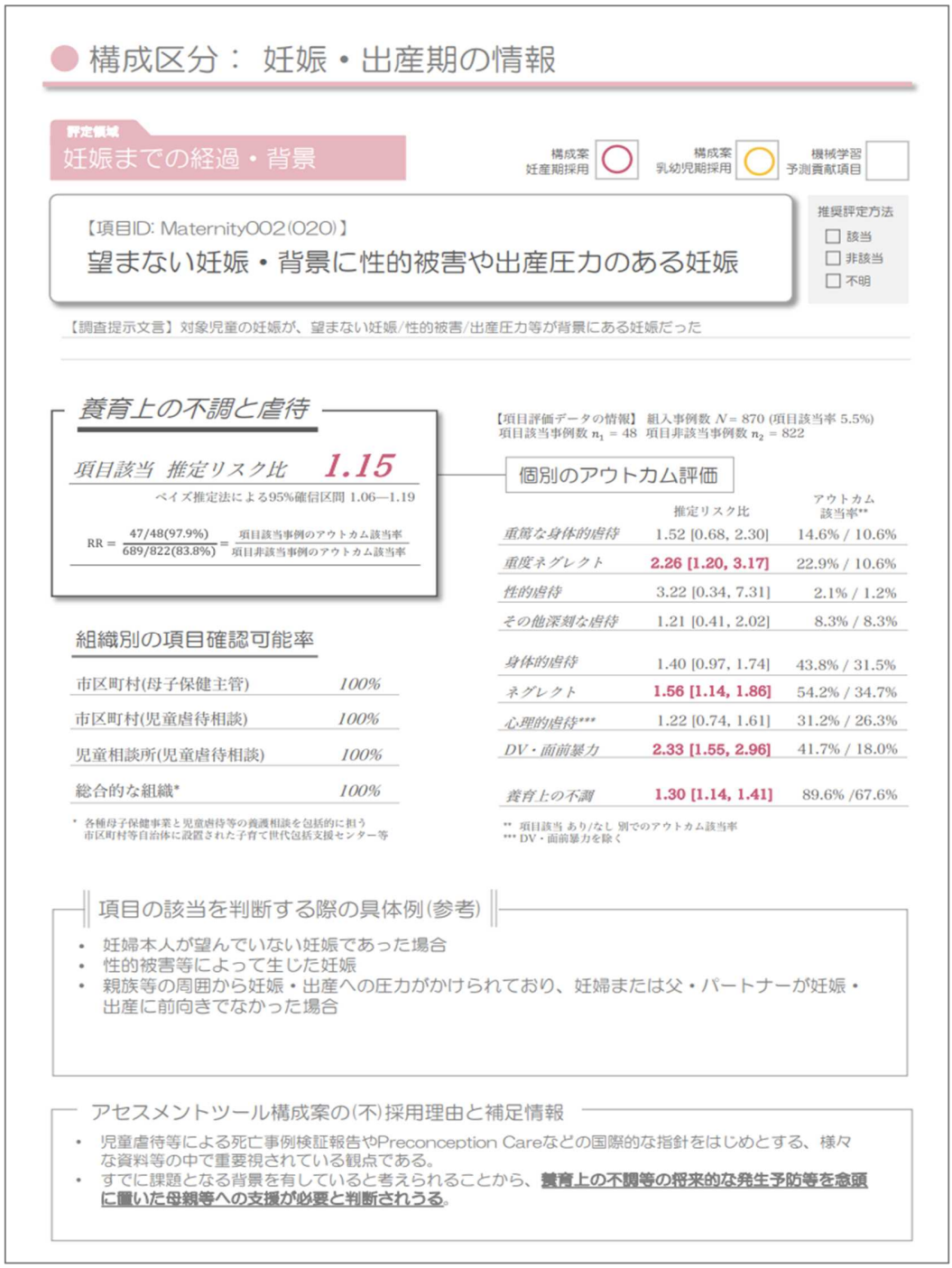



図 4.3 計画的欠損データデザインを用いた調査で得られた項目基礎情報の一例

※ 産業技術総合研究所(2022). 令和 3 年度 子ども・子育て支援推進調査研究事業「母子保健における児童虐待予防等のためのリスクアセスメントの在り方に関する調査研究」事業報告書サマリー& アセスメントツールの構成ガイドのアセスメント項目情報リスト p12 より引用(URL: https://staff.aist.go.jp/kota.takaoka/research/mhlw_parentChildHealth.html)

計画的欠損データデザイン等により、再犯・再非行に関する多面的なアセスメント項目の基礎情報を有しておくことは、高い精度での予測の実現や、各種 AI 倫理に関する妥当性を保証すること、あるいは妥当性について議論する上でも有意義であると考えられる。

### 4.2.4 CFP の反復測定(時系列情報)について

アセスメントは元来、単回の実施で終了するものではなく、支援や介入等の結果や状況の変化などに応じて随時見直される。CFP を用いたアセスメントにおいても、保護観察期間の開始時から、その経過に応じて、複数回の見直し・再アセスメントがなされて自然だと言える。

反復して測定される情報をリスク予測モデリングの枠組みで活用するにあたっては、データの組み入れ方について、いくつかのパターンが想定されうる。前提として、「常に最新の情報で予測結果を出力する」という、完全性を保証するための枠組みについては、どのような場合であっても共通するものと考えられる。

最も単純な形式は、CFP を用いたアセスメントを実施するタイミングを「初期」「中期」「後期」などの時期別で設定し、それぞれの段階でのアセスメントデータを用いてリスク予測モデルを構築するというものとなる。このとき、「初期と比べた中期の得点の変化量」など、時間的変化の情報を特徴量として学習に組み込むことなども想定される。

ここで、「リスク予測モデルの予測結果を受けて介入し、さらに予測を実施しようとする枠組み」における留意点を整理しておく。例えば、初期と中期の間などのタイミングで、「リスク予測モデルの結果を受けてどのような介入等を実施したか」によって、出力される予測結果は影響を受けうる。すなわち、「初回で再犯・再非行の高い予測値が出力されたがゆえに、重点的な介入が施される傾向が強くなり、その結果、再犯・再非行が生じにくくなった」というデータが蓄積されたならば、「本来的に高い確率での再犯・再非行が予測される事例であるにも拘らず、介入の効果を含んで再犯・再非行の低い予測値が出力されてしまう(そして重点的な介入の対象から外れてしまう)」といった事態が生じうる。ゆえに、各アセスメント段階までに発生した介入履歴等の情報は、考慮する対象として保有し、モデルに組み入れる工夫等を講じる必要があるものと考えられる。

反復測定される CFP アセスメントの経時的な情報は、「個々のタイミングで切り出し、過去の経過情報を組み入れて予測に用いる」といった形式以外にも、すべての時系列情報を一つのリスク予測モデルに学習させ、予測やシミュレーションに用いるといった形式でも活用可能性がある。ただし、CFP 等を含めた多面的なアセスメントツールを用いる場合、評価項目の数は通常複数になる(多変量)。また、事案の基礎情報などを組み合わせて利用することを想定すれば、それぞれの項目の尺度(名義尺度、順序尺度、間隔尺度、比率尺度など)や、とりうる値の範囲も異なりうるだろう(実数、0 より大きい実数、非負の整数、二値変数など)。そして、アセスメントが実施されうる回数には制限があり、また保護観察対象者によって異なることが想定される。アセスメントを実施するタイミングが不統一である場合には、不等間隔な時系列データとなる。さらに、前述の通り、アセスメントの実施前後に測定項目の変容を図った介入が施されることもあるだろう。このようなデータの生成過程や性質を想定した場合に、将来予測やシミュレーションを実現する単一のリスク予測モデルが適用可能であるかについては、個別に検証が必要となる点には十分に留意されたい。時系列情報を用いた評価や将来予測を実現しうる解析モデルの代表例には、時間的前後関係を考慮しながら項目間の相互関係を捉え、将来の状況変化に関する平均的な傾向(条件付き確率)を出力することができる動的ベイジアンネットワ

ークモデルや(e.g. Marcel et al., 2008)、時間的前後関係を構造的に取り入れたさまざまなニューラルネットワークモデル(Shterev et al., 2022; Bhanja, 2020)などがある。また、解釈性を重視し、理論的背景を持ってアセスメント情報等の時間的変化を予測する統計モデルの構築(e.g. Koslovsky et al., 2020)や、ネットワークアプローチに基づいて回顧的にアセスメント情報を解析し、個人や集団におけるアセスメント情報の変遷関係を捉えるモデル(e.g. Bringmann et al., 2013)などの手法活用も想定されうる。こういったモデルを活用する場合には、元来の目的に対して必要な出力情報を検討し、蓄積データの内容や形式を吟味した上で、目的とデータの性質に応じた最適な手法を選択する必要があるものと考えられる。

### 4.2.5 他の発展的可能性

ここまで、保護観察領域におけるリスク予測モデリングについて、CFP データを主軸とした再犯・再非行の予測を中心に、その技術的工夫の例について取り扱った。本節では、CFP や再犯・再非行という予測対象に必ずしも限定されない形式で、他の発展可能性について一例ずつ整理する。

第一に、再犯・再非行の防止を考える上で、その発生メカニズムや中核要因は、犯罪の種別等によって大きく異なりうるものと考えられる。例えば、物質依存や行為依存といった、再犯・再非行の可能性が精神医学的病理と密接に関わる犯罪種別もあれば、貧困や経済的自立等の側面を中核として、それに左右されうる犯罪種別も存在しうる。もちろん、犯罪種別のみで再犯・再非行の発生メカニズムが完全に切り分けられるということはなく、一つの事案に複数の要素が常時複合しているものと考えられる。しかし、例えば犯罪種別等でデータを切り分け、それぞれに対するリスク予測モデルを構築することには、いくつかの利点があるものと考えられる。区分別でモデルを切り分けるという発想によって、(1)特定の区分における予測性能が向上する(または予測精度が得られにくい区分がわかる)、(2)区分ごとに特化した研究の結果や区分に特化したアセスメント情報の組み入れと加味ができる、(3)項目貢献度指標等の数値が理論等に照らして解釈されやすくなる、(4)事案の背景ニーズに応じて異なる介入手法をデータ上整理しやすくなるなどである。犯罪種別等の区分によっては、例数が十分に得られず、モデル構築が困難となることは十分に想定されるが、今後の検討可能性がある方法論の一つとして特筆しておきたい。

第二に、保護観察領域におけるリスク予測モデリングの対象には、「再犯・再非行の発生」以外のものも十分に含まれうる。あくまで一例となるが、「再犯・再非行」と同程度あるいはそれ以上の水準に相当するものとして、「行方不明・失踪」や「死亡」など、最大限回避したい予防対象となるアウトカムなどの設定が考えられる。その他にも、「再犯・再非行の発生」の前段的要素となりうる個別の事象として、「対象者の家計の急変」などの経済的側面に関する事象や、「不良交友の再燃」などを中間防止指標として予測対象にすることなども検討されうる。

これらの発想は、個人データ利用に係る各種の原則(第五章参照)などを含めて十分に吟味する必要があるが、保護観察領域におけるアセスメントの強化や拡充を考えるにあたって、重要な評価の枠組みや、有意義な情報蓄積を検討する材料になるものと考えられる。

# 第五章　AI-CFP 実装に向けたロードマップ案の提示

本章では、前章までの議論を踏まえ、AI-CFP の実装に向けたロードマップ案の提示を行う。ロードマップは「令和 9 年度のリリース」を前提に作成し、対象とする範囲はリリース段階までの間に想定されることとする。システムのプロバイダ等の関与者の枠は設けず、本流となる検討事項とその時間的な流れを整理することを目標とする。そして、特に重要な基盤たる各種の原則や考え方の紹介など、今後の開発・実装を進めてゆく上での初期段階の検討事項に力点を置いた情報を整理する。

構成は、データの収集と取り扱い、実効的なサービスとして機能するシステム設計、AI の利活用に伴うリスクについて整理された主要な参照資源として、(1)個人データの取り扱いに係る「OECD 理事会勧告 8 原則」、(2) システムではなくユーザーが利用するサービスとして捉えた設計を志す「サービスデザイン」の考え方、(3)AI の利活用に伴って発生する各種のリスクを管理するフレームである「リスクチェーンモデル(RCModel)」(松本・江間, 2021; 東京大学未来ビジョン研究センター 技術ガバナンス研究ユニット AI ガバナンスプロジェクト, 2021 年 6 月)を紹介する。その他にも、本章には詳述しないが、AI の品質管理に関するガイドラインなども公開されてきている(国立研究開発法人産業技術総合研究所, 2022 年 7 月 14 日 https://www.digiarc.aist.go.jp/publication/aiqm/aiqm-referenceguide-v1.0-jp.pdf)。その上で、実装される AI を仮想的に設定し、想定されるロードマップの要素ごとに解説を加える。なお、本章で提示するロードマップ案やその概説内容は、事業受託者が一例として示す案である点に留意されたい。

## 5.1 主要な参照資源

### 5.1.1 OECD 理事会勧告 8 原則

データの利活用に係る大前提として、「個人データの取り扱い」に関する原則・指針・関連法制度については、必ず確認しておく必要がある。ここでは、OECD が 1980 年に公表し、2013 年に一部更新(update)のあった OECD Guidelines on the Protection of Privacy and Transborder Flows of Personal Data における 8 つの原則(BASIC PRINCIPLE OF NATIONAL APPLICATION, 以下、OECD8 原則と記載する)の概要を取り上げる。なお、当該原則は現在も更新に係る議論が展開されており(https://www.oecd.org/sti/ieconomy/privacy.htm)、最新の情報を参照されたい。ここでは、総務省による原則部分の邦訳(1980 年 9 月)を引用して紹介する。

**1．収集制限の原則**
個人データの収集には、制限を設けるべきであり、いかなる個人データも、適法かつ公正な手段によって、かつ適当な場合には、データ主体に知らしめ又は同意を得た上で、収集されるべきである。

**2．データ内容の原則**
個人データは、その利用目的に沿ったものであるべきであり、かつ利用目的に必要な範囲内で正確、完全であり最新なものに保たなければならない。

**３．目的明確化の原則**

個人データの収集目的は、収集時よりも遅くない時点において明確化されなければならず、その後のデータの利用は、当該収集目的の達成又は当該収集目的に矛盾しないでかつ、目的の変更毎に明確化された他の目的の達成に限定されるべきである。

**４．利用制限の原則**

個人データは、第９条（目的明確化の原則）により明確化された目的以外の目的のために開示利用その他の使用に供されるべきではないが、次の場合はこの限りではない。

(a)データ主体の同意がある場合、又は、

(b)法律の規定による場合

**５．安全保護の原則**

個人データは、その紛失もしくは不当なアクセス・破壊・使用・修正・開示等の危険に対し、合理的な安全保護措置により保護されなければならない。

**６．公開の原則**

個人データに係る開発、運用及び政策については、一般的な公開の政策が取られなければならない。

個人データの存在、性質及びその主要な利用目的とともにデータ管理者の識別、通常の住所をはっきりさせるための手段が容易に利用できなければならない。

**７．個人参加の原則**

個人は次の権利を有する。

(a) データ管理者が自己に関するデータを有しているか否かについて、データ管理者又はその他の者から確認を得ること。

(b)自己に関するデータを、i1)合理的な期間に、ii2)もし必要なら、過度にならない費用で、iii3)合理的な方法で、かつ、iv4)自己にわかりやすい形で自己に知らしめられること。

(c)上記(a)及び(b)の要求が拒否された場合には、その理由が与えられること及びそのような拒否に対して異議を申立てることができること。

(d)自己に関するデータに対して異議を申立てることができること及びその異議が認められた場合には、そのデータを消去、修正、完全化、補正させること。

**８．責任の原則**

データ管理者は、上記の諸原則を実施するための措置に従う責任を有する。

※ OECD Guidelines on the Protection of Privacy and Transborder Flows of Personal Data における PART TWO. BASIC PRINCIPLES OF NATIONAL APPLICATION について、
総務省による邦訳(1980 年 9 月)より引用。

原典:
https://www.oecd.org/sti/ieconomy/oecdguidelinesontheprotectionofprivacyandtransborderflowsofpersonaldata.htm

総務省による邦訳: https://www.soumu.go.jp/main_sosiki/gyoukan/kanri/oecd8198009.html

AI-CFP の実装(リスク予測モデルの実装)を検討するにあたり、これらの原則を遵守することは個人データの利活用という観点から必須事項であるものと考えられる。個人データを収集する上での同意や通知等の手続きについて、その妥当性がどのように担保されるか。あるいは、個人データをモデルに学習させ、新たな予測に用いるという枠組みが、個人データの(新たな)収集目的に合致するか。さらには、予測結果の利用範囲が明確化され、範囲外の利用が生じる恐れがあるかなど、事前に原則に照らした整理を行い、その透明性を確保する必要があるものと考えられる。

技術的な視点からは、保有・活用される個人データが常に最新でなければならないこと(データ内容の原則)より、再犯・再非行の予測を実行する際には、常に最新の情報が入力されるような設計が必要となる可能性がある。また、安全保護の原則と個人参加の原則より、入力となる個人データおよび予測結果は、セキュアな環境下で取り扱われ、出力時間等を含めたログ情報とともに再現可能な形で各種情報が保存される必要があるものと考えられる。

OECD8 原則は、リスク予測モデルの学習と利活用を実施する上で、特に(個人)データの取り扱いに関連する側面での重要な参照源となる。

### 5.1.2 サービスデザイン思考

提供者の視点ではなく、利用者の視点を中心にサービスや業務の設計を志す「サービスデザイン思考」の考え方について本章で紹介する。とくに、本事業に係るサービスは行政の一環として提供されることから、行政におけるサービスデザイン思考について取り扱っている資料(内閣官房 情報通信技術(IT)総合戦略室, 2018; サービスデザイン実践ガイドブック(β 版), 2018 年 3 月 : 以降「サービスデザイン実践ガイドブック」と呼ぶ)を参考にその内容を要約する。なお、参考資料は内閣府における標準ガイドライン群の一つの候補として作成された。

サービスデザイン実践ガイドブックでは、まず、民間企業においては利用者の体験等も考慮して設計された利便性の高いサービスが数多く存在している一方で、行政のサービスや業務は、必ずしも利用者視点で設計・開発がなされていないことが指摘されている。そこで、行政も利用者の視点を取り入れたサービスを打ち出し、サービスを通した利用者の目的達成を促進すべきであることが作成の背景・目的として示されている。

サービスデザイン思考とは、利用者を中心に物事を考える思考法をサービス設計に応用した概念とされる。より継続的にもしくはより多くの利用者にサービスを利用してもらい、その利用目的を達成してもらうためには、利用者がサービスに満足できることが不可欠であるとする思考法と言い換えることもできる。サービスデザイン思考で用いられる手法はいくつかあるが、一貫した考え方は、「サービス利用者の立場を考慮した調査・分析から得られる利用者の課題やニーズに基づき、サービスや業務を設計・開発する」ことである。

サービスデザイン実践ガイドブックでは利用者中心の行政サービスを提供するために必要となるノウハウが「サービス設計 12 箇条」としてまとめられている。各府省は、この考え方を踏まえてサービス・業務の見直しと抜本的な改革を進めるものとされている。以下に、サービスデザイン実践ガイドブックから「サービス設計 12 箇条」を引用して掲載する。

**第１条　利用者のニーズから出発する**

提供者の視点ではなく、利用者の立場に立って、何が必要なのかを考える。様々な利用者がいる場合には、それぞれの利用者像を想定し、様々な立場から検討する。サービス提供側の職員も重要な利用者として 考える。ニーズを把握するだけでなく、分析によって利用者が抱える課題・問題を浮き彫りにし、サービスの向上につなげる。

**第２条　事実を詳細に把握する**

実態の十分な分析を伴わない思い込みや仮説に基づいてサービスを設計するのではなく、現場では何が起きているのか、事実に基づいて細かな粒度で一つ一つ徹底的に実態を把握し、課題の可視化と因果関係の整理を行った上でサービスの検討に反映する。データに基づく定量的な分析も重要である。

**第３条　エンドツーエンドで考える**

利用者のニーズの分析に当たっては、個々のサービスや手続のみを切り取って検討するのではなく、サービスを受ける必要が生じた時からサービスの提供後まで（エンドツーエンド）の、他の行政機関や民間企業が担うサービスの利用まで含めた利用者の行動全体を一連の流れとして考える。

**第４条　全ての関係者に気を配る**

サービスは様々な関係者によって成り立っている。利用者だけでなく、全ての関係者についてどのような影響が発生するかを分析し、Win-Win を目指す。また、デジタル機器が使えない人も、IT を活用することによって便益を享受できるような仕組みを考える。

**第５条　サービスはシンプルにする**

利用者が容易に理解でき、かつ、容易に利用できるようにシンプルに設計する。初めて利用する人や IT に詳しくない人でも、複雑なマニュアルに頼らずとも、自力でサービスを利用して完結できる状態を目指す。また、行政が提供する情報や、利用者に提出や入力を求める情報は、真に必要なものに限定する。

**第６条　デジタル技術を活用し、サービスの価値を高める**

サービスには一貫してデジタル技術を用い、利用者が受ける便益を向上させる。技術の進展に対応するため、 IoT や AI 等の新技術の導入についても積極的に検討する。これまでデジタル以外の手段で提供してきたものであっても、業務の見直しによるデジタルへの移行の可能性を検討し、サービスの改善を図る。また、情報セキュリティとプライバシーの確保はサービスの価値を向上させるための手段であることを認識した上で、デジタル技術の活用によってサービスをセキュアに構築する。

**第７条　利用者の日常体験に溶け込む**

サービスの利用コストを低減し、より多くの場面で利用者にサービスを届けるために、既存の民間サービスに融合された形で行政サービスの提供を行うなど、利用者が日常的に多くの接点を持つサービスやプラットフォームとともに行政サービスが提供されるような設計を心掛ける。

**第８条　自分で作りすぎない**

サービスを一から自分で作るのではなく、既存の情報システムの再利用やそこで得られたノウハウの活用、クラウド等の民間サービスの利用を検討する。また、サービスによって実現したい状態

は、既存の民間サービスで達成できないか等、行政自らがサービスを作る必要性についても検討する。過剰な機能や独自技術の活用を避け、 API 連携等によってほかで利用されることを考慮し、共有できるものとするよう心掛ける。

**第９条　オープンにサービスを作る**

サービスの質を向上させるために、サービス設計時には利用者や関係者を検討に巻き込み、意見を取り入れる。検討経緯や決定理由、サービス開始後の提供状況や品質等の状況について、可能な限り公開する。

**第１０条　何度も繰り返す**

試行的にサービスの提供や業務を実施し、利用者や関係者からのフィードバックを踏まえてサービスの見直しを行うなど、何度も確認と改善のプロセスを繰り返しながら品質を向上させる。サービス開始後も、継続的に利用者や関係者からの意見を収集し、常に改善を図る。

**第１１条　一遍にやらず、一貫してやる**

困難なプロジェクトであればあるほど、全てを一度に実施しようとしてはならない。まずビジョンを明確にした上で、優先順位や実現可能性を考えて段階的に実施する。成功や失敗、それによる軌道修正を積み重ねながら一貫性をもって取り組む。

**第１２条　システムではなくサービスを作る**

サービスによって利用者が得る便益を第一に考え、実現手段であるシステム化に固執しない。全てを情報システムで実現するのではなく、必要に応じて人手によるサービス等を組み合わせることによって、最良のサービスを利用者に提供することが目的である。

※ 「内閣官房 情報通信技術(IT)総合戦略室, 2018; サービスデザイン実践ガイドブック(β 版), 2018 年 3 月」より引用。

保護観察における再係属のリスク予測モデルを実装する際も、上述の 12 箇条に照らした開発と設計を進めることが重要であると考えられる。利用者の満足度を向上させることは、サービスの利用率向上を促進し、ひいてはサービスを通した目的の達成(再係属率の低下等)に繋がると考えられる。

サービスデザイン思考では、利用者中心の視点で設計・開発していくことが重要であり、それを実現するための確立された手順・手法等があるわけではない。ここでは、サービスデザイン思考の典型的なモデル(手順の大枠)として知られるダブル・ダイヤモンド(Design Council, Framework for Innovation: Design Council's evolved Double Diamond)について紹介する。ダブル・ダイヤモンドでは、手順を課題の発見(図 5.1 左のダイヤモンド)と解決策の考案(図 5.1 右のダイヤモンド)の 2 つに大きく分け、さらにそれぞれが 2 つのフェーズから成るとして、以下 4 つのフェーズで整理されている。

1. **探索フェーズ(Discover)**
   想定される利用者等への調査を行い、利用者の視点から見た、本質的な課題やニーズを探索するフェーズ。ここでは制限を設けず、より広域的に課題を探索し、それらを列挙する。

2. **定義フェーズ(Define)**
   探索フェーズで列挙した課題群の優先度や実現可能性を考察し、課題の構造を理解する。それを基に、サービスを通して最終的に実現したい価値や目的を明確に定義する。探索フェーズで列挙した課題を小さく収束もしくは集約していくフェーズである。

3. **開発フェーズ(Develop)**
   定義した課題に対して、見込みのある複数の解決策を考案し、列挙していくフェーズである。ここでは制限を設けず、より広域的に解決策を展開していく。

4. **提供フェーズ(Deliver)**
   開発フェーズで提案された複数の解決策に対して試行や評価を繰り返す。これにより、解決策を洗練させ、最終的には一つのサービス・業務改善に落とし込む。

なお、各フェーズでのより具体的な手順や手法は、対象とするサービスや業務によって柔軟に組み換える、または、自ら考案していく必要があるとされている。具体例については、サービスデザイン実践ガイドブックも参照されたい。また、以上の 4 つのフェーズは必ずしも直線的・段階的に進むわけではなく、必要に応じて各ポイントで前のフェーズに立ち戻らなければならないこともある。
図. 5. 1 において、この様子は青色の点線矢印で表現されている。

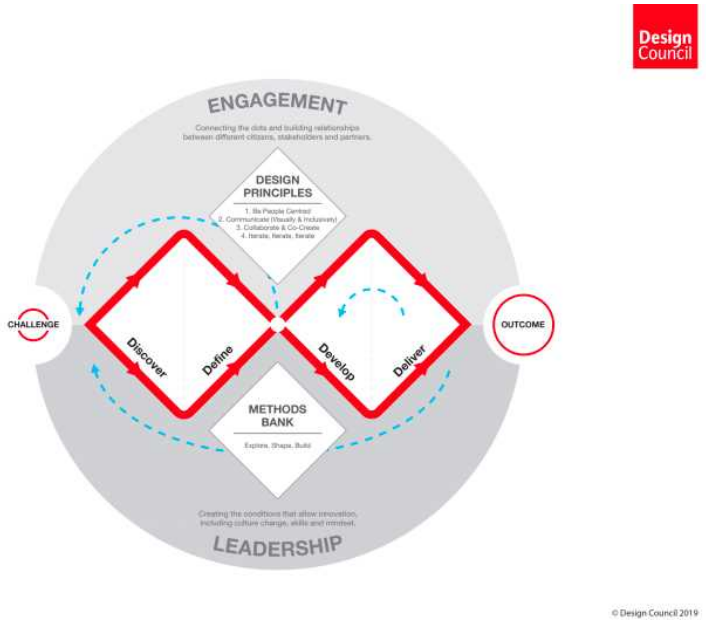



図 5.1 Design Council, Framework for Innovation: Design Council's evolved Double Diamond (https://www.designcouncil.org.uk/our-work/skills-learning/tools-frameworks/framework-for-innovation-design-councils-evolved-double-diamond/)

再犯・再非行のリスク予測モデルを実装する際の主な利用者は保護観察官である。あくまで一例ではあるが、リスク予測モデルのサービス実装を上記の 4 フェーズに照らし合わせて考えると以下のようになる。

| フェーズ | 内容 | 具体例 |
|---|---|---|
| 1. 探索 | 保護観察官を含めた実務を行う職員に対する調査を行い、再犯防止を図る上で重要な課題やニーズを現場視点で探索する。 | ● Web アンケート調査<br>● 現場でのヒアリング<br>● KA 法等を用いた上記結果の分析 |
| 2. 定義 | 優先度や実現可能性も踏まえて、再犯防止を図る上で重要な課題やニーズを絞り込む。必要であれば「再犯防止」という最終的な価値の定義も見直す。 | ● 利用者人物像の定義(ペルソナ設定)<br>● ジャーニーマップ等を用いた利用者体験の可視化<br>● ステークホルダー対象のワークショップ |
| 3. 開発 | 定義した課題に対して、AI を含めた ICT 技術による解決策、組織体制や業務の改革による解決策など、様々な解決の道筋を考案する。 | ● ストーリーボード等を用いたアイデア出し<br>● ステークホルダーによる議論と解決策の評価 |
| 4. 提供 | 上記で考案された解決策に対して試行と評価を繰り返す。最終的に一つのサービスへ絞り込み、洗練させる。 | ● (ICT 技術の場合)プロトタイプやモックアップの作成<br>● 利用者を交えた試行と評価<br>● 運用テスト<br>● リスクの評価<br>● 倫理面の精査 |

＜別添 1　サービス設計 12 箇条とそれぞれの確認ポイント＞
https://cio.go.jp/sites/default/files/uploads/documents/servicedesign_betten1.pdf

### 5.1.3 AI のリスクチェーンモデル

本章では、AI の利活用に伴って発生する各種リスクを管理するフレームである「リスクチェーンモデル(RCModel)」(松本・江間, 2021; 東京大学未来ビジョン研究センター 技術ガバナンス研究ユニット AI ガバナンスプロジェクト, 2021 年 6 月)について各種資料(https://ifi.u-tokyo.ac.jp/projects/ai-service-and-risk-coordination/)を参考に要約する。
RCModel の主な目的は二つあり、 1)重要なリスクシナリオの識別、2)識別されたシナリオに対するリスク軽減策とリスクマネジメント計画の立案である。このフレームワークでは、登場人物を(1)AI モデル及び AI システムの開発者、(2)サービス提供者、(3) ユーザーの 3 つに分類しており、最終的に誰がどのリスクに対応するかの役割分担が明確になる。具体的な方法は、図 5.2 に示すように、5 つの

STEP で構成されている。主に Step1 と 2 が目的の 1)に対応しており、Step3 から 4 が目的の 2)に対応している。図 5.2 に付け加える形でそれぞれの Step の注意点を以下に述べる。

＜Step1 の注意点＞
「人間よりも高度な検知」等のビジネス目的だけでなく「不適切な利用の防止」等の社会的責任も同時に定義する。後のリスク検討を容易にするため、複数の価値・目的がある場合はそれらに優先順位をつける。(Step1 のイメージ：図 5.2 左上)

＜Step2 の注意点＞
Step1 で定義した価値・目的の優先順位も考慮した上で、リスクシナリオごとの優先度を検討する。(Step2 のイメージ：図 5.3 左下)

＜Step3 の注意点＞
リスクチェーンの起点になる要因、矢印の方向(方向の有無も含めて)は扱うリスクによって変化する。(Step3 のイメージ：図 5.3 右上)

＜Step4 の注意点＞
リスクの大きさと費用対効果を踏まえて最適なリスクコントロールを検討する。先に述べた登場人物の分類ごとに、とるべき対策や責任の範囲を明確にしておく。

＜Step5 の注意点＞
リスクシナリオごとに各ステークホルダーの役割を整理し、関係者間で共通認識を持つ。(Step5 のイメージ：図 5.3 右下)

実際に再犯予測の AI を導入する際は、ここで紹介したようなフレームワーク等を用いてリスク管理することが、1.4.1 章で述べた検討課題の観点からも重要になる。また、リスクシナリオの検討では、一般の AI システムに共通のリスクに加えて、再犯予測 AI に特有のリスクについて有識者を交えて議論することが必須であると考えられる。


図 5.2 RCModel におけるケース検討のステップ。スライド「リスクチェーンモデル(RCModel)ケース検討事例：Case06 再犯可能性の検証 AI，東京大学未来ビジョン研究センター 技術ガバナンス研究ユニット AI ガバナンスプロジェクト(https://ifi.u-tokyo.ac.jp/projects/ai-service-and-risk-coordination/)」p.2 からの引用。


図5.3 RCModel における各Step のイメージ図。スライド「リスクチェーンモデル(RCModel)ケース検討事例：Case06 再犯可能性の検証 AI，東京大学未来ビジョン研究センター 技術ガバナンス研究ユニット AI

ガバナンスプロジェクト(https://ifi.u-tokyo.ac.jp/projects/ai-service-and-risk-coordination/)」p.5,9,10,13を抜粋して引用。

※ RCModel は、東京大学未来ビジョン研究センター技術ガバナンス研究ユニットの AI ガバナンスプロジェクトによって開発されたフレームワークであり、当該モデルはクリエイティブ・コモンズ・ライセンス CC-BY4.0 の下で公開されている。モデルの実施や共同研究を検討する場合は、東京大学未来ビジョン研究センター 技術ガバナンス研究ユニット事務局（ifi_tg@ifi.u-tokyo.ac.jp)への連絡が必要となる。（東京大学未来ビジョン研究センター 技術ガバナンス研究ユニット AI ガバナンスプロジェクト，2021 年 6 月より）

## 5.2 例として用いる「仮想 AI」

ここまで、個人データと機械学習(AI)技術を駆使したリスク予測モデルについて、その主要な参照資源を紹介した。本節以降は、保護観察の文脈における AI-CFP の実装を具体的に念頭においた、実装までの概略的な検討手順をロードマップの形式で整理する。

ロードマップの設定と個別トピックに関する議論を実施するにあたり、本章の範囲内で用いる「仮想的なリスク予測モデルの形」を暫定で次の通り設定する(以下、仮想条件)。なお、設定された各種の条件は、ロードマップの解説を了解しやすくするため暫定的に設定したものであり、理想的な条件等を示すものではないことに留意されたい。

表 5.1 本章で設定する仮想条件の概要

| 区分 | 個別観点 | 内容 |
|---|---|---|
| 目的 | モデルが果たす機能と価値 | 【目的】再犯・再非行に関するアセスメントの補助・拡充に寄与する情報の提供<br><br>【価値】高度化されたアセスメント情報に基づく処遇等の展開による再犯・再非行の発生防止 |
| ユースケース | 具体的場面 | 保護観察開始時または処遇方針見直し時における CFP を用いたアセスメント実施のタイミング |
| | 活用方法 | 出力情報を処遇決定の参考情報として閲覧する |
| モデル概要 | 予測対象(アウトカム) | 5 年以内の再犯・再非行 |

| | | |
|---|---|---|
| | 出力内容(主) | アウトカムの蓋然性に関する予測値(確率やスコア等) |
| | 出力内容(副) | 出力内容(主)で得られた数値の解釈性や処遇決定への活用をサポートする情報(Shapley Additive Explanations や Counterfactual Explanations など) |
| | 入力情報 | ・事案の基礎データ(年齢・性別等)<br>・CFP アセスメントデータ |
| | 基本的性能 | 既存のアプローチ(組織決定アセスメントアプローチ/保険数理的アプローチ)よりも高い予測性能が得られている |
| システムとハードウェア | システム | 業務インフラとなっている情報管理システムへ機能を追加・搭載。解析を実行するサーバ等を、システムを動作させるサーバとは別に設置して中間 API を用いて連結・動作させる |
| | 端末 | 業務用の PC 端末からサーバにアクセス |

## 5.3 想定される基本の流れ(ロードマップ案の概念図)

前節に示した仮想条件のモデルを実装することを想定した場合、一例として、図 5.4 に示すような導入までの流れが想定される。

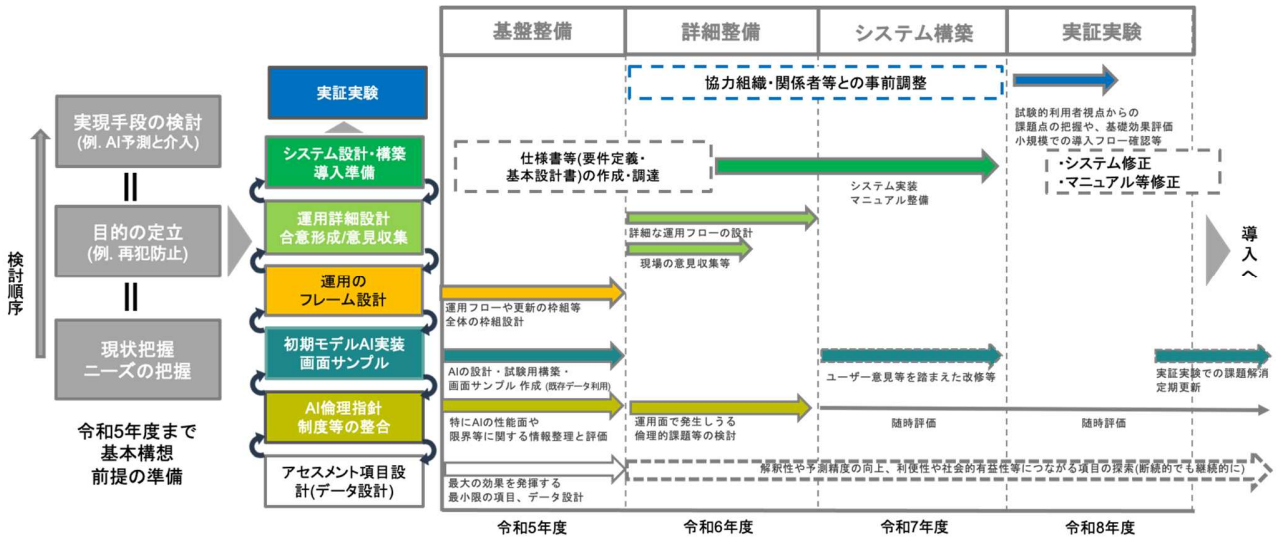



図 5.4 AI-CFP の導入に係る開発等のロードマップ案【概念図】

図 5.4 では、「令和 9 年度の導入」を前提とした時の、主要な検討課題のコンポーネントと検討フェーズが示されている。大きなフェーズとしては、最も左にある「前提の準備」の段階、「基盤整備」の段階、「詳細整備」の段階、「システム構築」の段階、「実証実験」の段階を設定し、それら

を経た上での「導入」が図示されている。検討の枠組みとなる各コンポーネントについては、下部から順に「アセスメント項目設計」「AI 倫理指針・各種制度等との整合」「初期モデル実装/画面サンプルの作成」「運用のフレーム設計」「運用詳細設計、合意形成・意見収集」「システム設計構築・導入準備」「実証実験」で構成されている。また、それらの前段として「現状把握・ニーズの把握」と「目的の定立」「実現手段の検討」の対応関係を整えることが、それぞれ事前の準備事項として位置付けられている。以降では、各コンポーネントの内容を概説する。

### 5.3.1 現状とニーズの把握・目的の定立・実現手段の検討

第一に、AI-CFP 導入の前提として、(潜在的な)ニーズから全てをスタートする必要がある。最も根幹かつ重要なフェーズであると形容されるだろう。ニーズに合致しない不要な機能は、必然、実装されても利用されない。「サービスデザイン思考」でも扱われる通り、サービス(システム)を利用するユーザーや、その利用によって恩恵を受ける対象者等の視座から、「何が必要とされているか」を十分に調査・吟味する必要がある。これは、単に提供者側の想像や思い込みではなく、ユーザー側のリアルな意見情報や調査実態データ等に基づき、一般化可能な水準で最大限具体的に整理されることが望ましい。

現状やニーズが把握されたら、「解決したい課題と目的の定立」を実施する。このとき、設定される目的や解決したい課題の設定は、提供されるサービスの機能を前提とせず、純粋に、明らかとなったニーズから展開される必要がある。

目的が明確に定まれば、実現のための手段を検討する。このとき、必ずしも「AI」やリスク予測モデリングが、(1)当該目的を達成するための必要な手段であるとは限らない、(2)優先的に選択される手段であるとは限らない、(3)目的達成までのステップ等を想定した時に最初に着手する手段であるとは限らない、といった点に留意されたい。また、サービスの設計者は、「選択しようとしている手段が、どのような性質を持つ道具であるか」についても、十分に理解しておく必要がある。

本節の作業は、リスク予測モデリングの「基本的構想」を定式化するものとなる。以降は、リスク予測モデリングが「ニーズに合致した目的の達成にふさわしい手段である」ことが確認されたという前提で、個別の検討観点について整理する。

### 5.3.2 アセスメント項目設計(データ設計)

図 5.4 には、基盤整備段階から導入後以降にかけて、アセスメント項目設計(データ設計)が継続的に検討される必要があることを示している。

機械学習技術等を用いたリスク予測モデリングを実現する上で、モデルの選択やシステムの調整など、検討する観点は複数存在する。しかし、その基盤的意味において、最も大きな役割を担うのが「データの設計」であると考えられる。

一例として、「再犯・再非行の予測」を目的とした場合、その本質を的確に捉える項目情報があれば、高度で複雑な機械学習モデルは不要となる可能性もある。少数の本質的な項目情報に絞ることができれば、設計はシンプルになり、ユーザーによる情報入力の負担は軽減され、結果的に質の高いデータが蓄積され、システムや運用・保守の負荷は小さくなり、得られた予測結果が解釈・整理しやすく、開発・更新のための他のデータ項目を追加検討する余地なども生まれる。もちろん、「再犯・再非行」が発生するメカニズムは多様であり、少数の項目だけでその全てが記述されうる訳ではないも

のと考えられるが、ここでは、データの設計が全体の機能性に大きく関与するということを共有しておきたい。

リスク予測モデルに学習させ、実行時に入力する情報には、さまざまなものが想定されうる。本稿の第二章および第三章で取り扱った、基礎情報やCFPアセスメント情報などが一例となる。このとき、「再犯・再非行」の予測を考えるにあたって、特にその原因やリスク要因となりうる事柄についてのアセスメント結果は、少数の項目で最大限の予測的妥当性を得てゆく上での主力的な情報源となりうる。再犯・再非行に係る重要なアセスメント観点について研究し、整理することは、リスク予測モデリングとは独立して「アセスメントの強化・拡充」のためにも重要な取組となる。

リスク予測モデリングにも活用するアセスメント項目等を事前に研究し、設計しておくことは、その他にもさまざまな利点を生み出すものと考えられる。具体的には、(1)項目の該当頻度や評定者内・間の信頼性、介入による反応性などの基礎情報があらかじめ把握された項目を利用することができる、(2)社会的バイアスが混入する可能性があらかじめ評価でき、そのリスクを最大限排除した項目セットを選択することができる、(3)測定範囲の代表性について検討可能な資料が得られるなど、リスク予測モデリングの実装にまつわる多くの倫理的側面について、評価に必要な情報が的確に準備できるといった側面もある。

アセスメント項目等を効率的に探索し、選抜するための調査方法の一例は、4.2.2節に示した計画的欠損データデザインを参照されたい。

### 5.3.3 AI倫理指針・各種制度等との整合

図5.4には、特に基盤整備と詳細整備のフェーズで検討の力点が置かれつつ、導入後にかけて随時評価される形式で、「AI倫理指針・各種制度等との整合」に関する検討の必要性が示されている。なお、倫理的側面に不備がある技術は、他の側面がいかに優れていたとしても、導入そのものの是非が再考されることとなりうる。サービスの根幹に係る検討事項であると言えるだろう。

倫理的側面の評価や、各種制度等との整合関係については、第一章に示したような一般的な評価観点や、本章に紹介したいくつかの原則やフレームワークなど、いくつか基本的な参照資源はあるものの、ユースケースによって個別に検討すべき観点は異なってくる。個人データ活用の是非、倫理的側面の評価、関連する規則や制度との整合、それらを遵守するための運用上の取り決めなど、多面的な眼差しをもって評価・検討する(継続的な)機会を設ける必要があると考えられる。

なお、リスク予測モデルに関する倫理的評価の観点の中には、「データ自体の評価」や「モデルの性能評価」など、データを用いて定量的に評価が必要な観点も存在する。したがって、次節に整理するリスク予測モデル(初期モデル)の構築手続きと並行して進めてゆく必要がある。

### 5.3.4 初期モデルAI実装・画面サンプルの作成

設計されたデータが一定以上蓄積されている場合、リスク予測モデルを実際に構築し、画面上で動作するサンプルを作成することには有用性がある。「これからどのようなものを作り、運用しようとしているのか」について、具体的なイメージの共有が可能となる。それにより、必要・不要な機能の洗い出しや、運用上起こりうる誤用のパターン想定など、実装に向けたさまざまな議論を後押しすることにつながる。また、技術的側面からは、出力情報の内容や提示の方法、動作の重量感などを確認することができ、さまざまなエラー等の可能性に対して対策を講じる機会の確保につながる。

リスク予測モデルの構築手続き自体は、第二章に実践的に示した手続きによって、その骨格自体は実現される(なお、第二章に示した手続きは試験的解析を目的としたものであり、全体としては検証手続きの内容にも評価事項にも不足がある)。これに加えて、学習データ自体の基礎的な評価や、複数モデルでの性能の比較検討、その他、より解釈性や活用可能性を高める補足的な解析の適用など、目的に応じて必要な解析手続きや評価観点は変化する。当該手続きによって、性能等の資料の他、「学習済みモデル」と「それを動作させるためのプログラム」を得ることとなる。

議論の共通土台としての、学習済みモデルを動作させる画面サンプルの構築方法にもいくつかの手段がある。(1)簡易的なWebアプリケーションを作成し特定の端末やネットワーク内で動作させる方法や、(2)本番と同等の仮想的な環境を構築してサンプルを共有する方法などである。モデルの具体的な動作の特徴や癖などを確認する必要がない場合には、モデルを動作させるアプリケーション等の画面と画面遷移の関係等を構造化するソフトウェア等により、疑似的なアプリ画面を構成して共有するなどの手段も考えられる。

### 5.3.5 運用のフレーム設計

図5.4では、「運用の基本的なフレームの設計」を基盤整備のフェーズでの必要検討事項に位置付けた。

ミクロには、誰が、いつ、どのようなタイミングで、どのような端末から情報を入力し、どのタイミングで出力結果を得て、どのような場でそれを参照し、結果を誰とどのように判断し、どのような意思決定に用いて、その結果どのような効果が期待されるのかという一連の流れを設定する作業となる。また、マクロには、リスク予測モデルの活用に必要な知識と使い方をどのように定義し、それをどのようにどの程度の頻度で伝え、誰がサービス全体を維持・管理・監督し、導入の効果をどのように評価するのかなどの管理・マネジメントに関する枠組みを定める作業となる。このとき、リスク予測モデルの導入対象となる組織によって、業務の流れが全く異なっている場合なども十分に想定されうる。そのような場合、リスク予測モデルの運用に関する設計が馴染まず、結果的に利用することができなくなるといった事態の発生も想定されうる。このような場合には、運用の設計を見直すか、場合によっては業務の流れを標準的に統一するなどの取り組みが合わせて必要となる可能性も想定される。

また、機械学習技術は、道具であり、さまざまな限界点が内包されている。道具の用途や限界に由来する各種リスクや「誤った活用方法」などについては、ユーザーがそれを十分に理解し、適切に活用する必要がある。リスク予測モデルを適切に活用するための運用の枠組みには、技術の内容や使い方等に関する研修など、継続的に活用を支えるための取り組みが含まれるという点にも留意されたい。

### 5.3.6 運用詳細設計・合意形成・意見収集

図5.4では、詳細整備のフェーズに「運用詳細設計・合意形成・意見収集」の検討を設けた。これは、前節の「運用フレームの設計」と連続する内容となるが、ヒヤリング等による現場意見の収集と反映に力点をおいた、運用フレームの調整的側面に主眼がある。ここまでに設計してきた目的やサービスの内容などを詳細に共有し、実践的に活用可能な形式をユーザーとともに模索する段階と位置付

けられる。これに加えて、導入に向けた合意形成の下地を整えるフェーズとしても機能することが想定される。

ユーザーによる情報入力などの負担や、活用に一定の理解を必要とするサービスを導入する場合には、その維持と浸透を図る上で「事前の合意形成」は重要な役割を担うと考えられる。そして、リスク予測モデルの目的や適切な理解を共有しながらユーザーとともに導入を進めてゆく上で、「ユーザーの意見が反映される、ユーザーとともにサービスを創ってゆく」という枠組みを設けておくことは、導入やその後のメンテナンスに係る合意形成を円滑に進める上でも有効に機能しうるだろう。

ユーザーの合意が得られるモデルの設計やユースケースの詳細が明らかになれば、リスク予測モデルを搭載するシステムの詳細な要件が決定できる。また、この段階で実現困難なニーズがある場合、「それを将来的にどのように実現するか」を想定することが可能となる。新規のデータ蓄積や新機能の追加予定など、将来の更新可能性を念頭に組み入れて、システム構成に柔軟性を与えたり、次期のモデル更新を見据えた入力データ項目の組み入れを講じたりすることもできるだろう。

### 5.3.7 システム設計・構築・導入準備

保護観察領域におけるリスク予測モデリングの実装を例として考える場合、システム設計の詳細をここで十分に議論することは叶わない。リスク予測モデルの実装媒体を、現在業務で利用している基幹システムに搭載するのか、それと連動する別のシステムを構築するのか、あるいは全く独立した環境にそれを構築するのかといった条件に依存して、想定される手続き等が全く異なりうるためである。しかし、どのような場合であっても、その要件は前節までに整理してきたユーザーのニーズ等を十分に反映し、実現しているものである必要があることは言うまでもない。画面デザイン(UI)、リスク予測モデルを動作させるサーバ側の設計、システム全体を通じて得られるユーザー体験(UX)の評価など、システム開発の基本的事項等をおさえた開発を実施することになる。

また、リスク予測モデルを「サービス」として捉えた場合、円滑な導入と維持に必要な各種の準備も必要となる。一例をあげれば、サービスマニュアルの作成や、リスク予測モデルを用いたアセスメントの実践に関する研修のパッケージなどである。次節に述べる「実証実験」の実施を目標としながら、必要な準備を整えてゆくのが実際的であると思われる。

### 5.3.8 実証実験

新規に実現しようとするアイデアやサービスが有効に機能しうるかどうかを、コンパクトな形で試験的に検証する枠組みとして、Proof of Concept(PoC)を行う場合がある。本節では、各種要件を満たして構築された導入前提のリスク予測モデルについて、その最終テストを行う枠組みとして「実証実験」を位置付け、その実施を提案する。導入によって影響が及ぶ範囲が極めて大きいサービスの場合に、本番同様の形式で導入前テストを一部の範囲に絞って行うという発想は自然であるだろう。なお、サービスを実際に導入した後になって、初めて明らかになる課題も多い。サービスに対する問い合わせの頻度やその内容、本格導入後に想定される諸事項への対応に係る体制規模などについても、実証実験によって部分的に明らかにすることができる。データの入力と蓄積と予測結果の参照が循環する枠組みにボトルネックが発生し、エコシステムが機能しないといった事態の発生も想定されうる。サービス提供者側が、各種課題や対応の流れを本番導入前に経験し、把握する機会としての意義も大きいだろう。

実証実験では、サービス全体や、UI/UX などの個別の観点等について、その課題や利便性等に関する評価情報を得るために、アンケート調査等を事前に設計しておくことが望ましい。さらに、評価項目や評価方法は事前に公表され、調査等がなされた全ての結果について完全に報告される必要がある。当該評価手続きは、サービスの利点や課題を明らかにし、建設的な改善に示唆を与え、透明性を担保し、そしてリスクと有益性のバランスを総合評価する上での重要な資料となる。

# 第六章　総括

本調査研究事業の実施背景となった大元の目的は、(1)特に長期の再犯・再非行リスクに影響する要因についてデータ解析を行い、リスク予測モデリングに必要な項目や特徴量について基礎的な知見を洗い出すこと、(2)今後、長期再犯・再非行のリスクおよび急性再犯・再非行リスクの予測等に必要な開発・実装を行い、令和9年度頃の運用を目指すこととした場合の具体的課題を整理し、(CFP等のアセスメントを用いた)AI導入に向けた当面の具体的なロードマップについて検討することであった。

当該目的を達成するための前段として、第一章では、保護観察領域における再犯・再非行防止のためのアセスメントの基本的発想(Risk-Need-Responsivityモデル)を踏まえた上で、CFP等のアセスメントツールの発展的な利用形式としてリスク予測モデリングを位置付け、その利点や限界点を整理した。また、機械学習技術等の活用を前提としたリスク予測モデリングの導入を検討するにあたっての、開発指針や倫理的側面に関連する検討事項と、実装から導入にかけての個別検討課題について、前提となる大きな枠組みを共有した。これにより、本事業の範囲で取り扱われるデータ解析による基礎検証が、AI-CFPの開発と導入におけるどの個別課題について検討するものかを明確にした。リスク予測モデリングの開発と導入に係る唯一完全なガイドライン等は存在せず、構想するモデルやユースケースによって必要な評価観点や遵守すべき事項は異なる。今後の開発と導入に向けた各種取り組みの基礎となるという点で、必要な評価観点が多面的に提案されたことの意義は大きいと考えられる。

保険数理的アプローチを超えて、リスク予測モデリングの導入を試みることの中核的意義は、精確な予測の実現にある。第二章では、事件管理システムに蓄積された事案の基礎データを用いて、5年以内の再犯・再非行の発生をどの程度の精度で予測可能かについて、予測的妥当性の基礎評価を実施した。その結果、異なるデータや条件間で単純比較は困難であるものの、長期的な再犯・再非行の予測が基礎データだけであっても一定の精度で実現できることが示された(AUC-ROCs ≧ 0.713)。ただし、より精確な予測を目指す視点や、予測結果を解釈し各種処遇等の検討に活用してゆく上では、基礎データのみを用いたリスク予測モデルの実装には不足があると考えられた。

再犯・再非行に対する予測精度の向上や、より実践的に結果解釈し、それを活用するためには、対象者のアセスメントデータを活用する必要がある。第三章では、データの蓄積期間に限りがあるものの、CFPアセスメント情報と短期的な再係属との関連についての基礎的な解析を実施した。その結果、CFPデータを活用することでも、短期的な再係属を一定以上の精度で予測できることが示された(AUC-ROCs ≧ 0.692)。当該結果は、既存の保険数理的アプローチよりも予測精度の視点で優れていると評価されうる。しかし、総合的な視座からして、既に十分な性能が得られていると形容されるものではない。予測的妥当性の向上を含めた今後の発展を考える上で、重要なアセスメント項目を評価・選抜し、項目化するなどの手法が功を奏する可能性が指摘された。

第四章では、第二章および第三章の解析結果を踏まえ、今後の開発に寄与しうるいくつかの技術的工夫の例を紹介した。実践的に結果を解釈し、各種処遇等への援用に貢献しうる「反実仮想説明」の概要や、予測的妥当性や解釈可能性の向上のみならず、リスク予測モデリング全体の基礎を支えるアセスメント項目の設計を補助する計画的欠損データデザインに基づく調査法などである。ロードマップに示す具体的な検討課題について、課題解決のための提案手段の一つとなっている。

第五章では、AI-CFPの開発と導入に向けて、主要な参照源となりうる個人データの取り扱いに関する原則、サービスデザイン思考、AIのリスク管理に関するフレームワークを紹介した。その上で、令

和 9 年度頃以降の導入を前提としたロードマップ案を作成した。ロードマップは、根幹となる基本的構想の定式化に相当する「現状把握・ニーズの把握」と「目的の定立」と「実現手段の検討」の対応関係が整っていることを前提に、(1)アセスメント項目設計、(2)AI 倫理指針・各種精度等との整合、(3)初期モデル実装/画面サンプルの作成、(4)運用のフレーム設計、(5)運用詳細設計・合意形成・意見収集、(6)システム設計構築・導入準備、(7)実証実験のコンポーネントで構成された。直近の検討観点に位置付けられた(1)～(4)は、リスク予測モデリング導入そのものの是非に関わる事項を含むことや、今後の積み重ねに大きく影響しうるという点で、重要であることが整理された。

# 引用文献・参考文献

＜第一章＞

法務総合研究所(2021). 令和 3 年版犯罪白書ー 詐欺事犯者の実態と処遇 ー
URL: https://www.moj.go.jp/content/001365724.pdf

Bonta, J., & Andrews, D. A. (2017). The psychology of criminal conduct. 6th ed. New York, NY; Routledge.

勝田聡・羽間京子(2020). 保護観察における新たなアセスメントツールー期待される効果と課題－千葉大学 教育学部研究紀要, 68, 317-322

羽間京子・勝田聡(2021). 保護観察におけるアセスメントツールの再犯予測力の検証, 千葉大学 教育学部研究紀要, 69, 27-32.

Andrews, D. A., Bonta, J., Wormith, J. S. (2004). The Level of Service/Case Management Inventory (LS/CMI). Toronto, Canada: Multi-Health Systems.

Andrews. D. A., Bonta, J., Wormith, J. S. (2006). The recent past and near future of risk and/or need assessment. Crime & Delinquency, 52, 7-27. doi: 10.1177/0011128705281756

P. Gillingham (2016). Predictive Risk Modeling to Prevent Child Maltreatment and Other Adverse Outcome for Service Users: Inside the 'Black Box' Machine Learning.

Mickelson, N., Laliberte, T., & Piesher, K. (2017). Assessing Risk: A Comparison of Tools for Child Welfare Practice with Indigenous Families, Center for Advanced Studies in Child Welfare, University of Minnesota.

Marshall, D. B., & English, D. J. (2000). Neural network modeling of risk assessment in child protective services. Psychological Methods, 5, 102-124.

Russell, J, (2015). Predictive analytics and child protection: Constraints and opportunities. Child Abuse & Neglect, 46, 182-189.

Vaithianathan, R., Maloney, T., jiang, N., Dare, T., de Haan, I., Dale, C., & Putnam-Hornstein, E. (2012). Vulnerable children: Can administrative data be used to identify children at risk of adverse outcomes? Auckland, New Zealand: University of Auckland.

Baumann, D. J., Law, J. R., Sheets, J., Reid, G., & Graham, J. C. (2005). Evaluating the effectiveness of actuarial risk assessment models. Children and Youth Services Review, 27, 465-490.

Crea, T. M. (2010). Balanced decision making in child welfare: Structured processes informed by multiple perspectives. Administration in Social Work, 34, 196-212.

Child Welfare League of America (2005). A comparison of approaches to risk assessment in child protection and brief summary of issues identified from research on assessment in related fields.

Gillingham, P., & Humphreys, C. (2010). Child protection practitioners and decision-making tools: Observations and reflections from the front line. British Journal of Social Work, 40, 2598-2616.

髙岡・坂本・橋本・北條・鈴木・菊池・古川他(2020). 児童虐待対応におけるアセスメントの在り方に関する調査研究 ー系統的な項目収集・全国横断 Web 調査によるアセスメント項目の基礎評価と研究知見に基づく市区町村および児童相談所で利用可能なセーフティアセスメントツール案の構成ー, 令和元年度 厚生労働省 子ども・子育て支援推進調査研究事業, 調査研究報告書. URL: https://staff.aist.go.jp/kota.takaoka/Ai%20for%20better%20society_files/pdf/2020project20-report.pdf

髙岡・北條・山本・難波・椎名・飛澤・柳他(2021). 児童虐待対応におけるアセスメントの在り方に関する調査研究ー (a) 児童相談所および市区町村で実践的に活用可能なセーフティアセスメントツールの開発と予測的妥当性・評定者間一致性の検証 (b) アセスメントツールの活用実態と今後の活用の在り方についてー, 令和 2 年度 厚生労働省 子ども・子育て支援推進調査研究事業, 調査研究報告書.

Vaithianathan, R., Putnam-Hornstein, E. , Jiang, N., Nand, P., & Maloney, T.(2017, April). Developing Predictive Models to Support Child Maltreatment Hotline Screening Decisions: Allegheny County Methodology and Implementation.

Shlonsky, A., Wagner, D. (2005). The next step: Integrating actuarial risk assessment and clinical judgment into an evidence-based practice framework in CPS case management, Children and youth services review,27,409-427.

Liao, S. H., Chu, P. H., & Hsiao, P. Y. (2012). Data mining techniques and applications - A decade review from 2000 to 2011. Expert Systems with Applications, 39, 11303-11311.

Tsai, H. H. (2012). Global data mining: An empirical study of current trends, future forecasts and technology diffusions. Expert Systems with Applications, 39, 8172-8181.

髙岡・坂本・古川・椎名・緒方・遠藤・山本・柳・坂上(2022). 令和３年度 子ども・子育て支援推進調査研究事業「母子保健における児童虐待予防等のためのリスクアセスメントの在り方に 関する調査研究」調査事業報告書, URL: https://staff.aist.go.jp/kota.takaoka/Ai%20for%20better%20society_files/pdf/2021project33-report.pdf

Dare T (2015) The ethics of predictive risk modeling. Challenging child protection: New directions in safeguarding children: 64-76.

Keddell, E (2014) Current debates on variability in child welfare decision-making: A selected literature review. Social Sciences 3: 916-940.

Blank A, Cram F, Dare T, de Haan I, Smith B, Vaithianathan R (2015) Ethical issues for Māori in predictive risk modeling to identify new-born children who are at high risk of future maltreatment.

Gillingham P (2006) Risk assessment in child protection: Problem rather than solution? Australian Social Work 59: 86-98.

Braverman DW, Doernberg SN, Runge CP, Howard DS (2016) OxRec model for assessing risk of recidivism: ethics. The Lancet Psychiatry 3: 808-809.

Keddell E (2015) The ethics of predictive risk modelling in the Aotearoa/New Zealand child welfare context: Child abuse prevention or neo-liberal tool? Critical Social Policy 35: 69-88.

de Haan I, Connolly M (2014) Another Pandora's box? Some pros and cons of predictive risk modeling. Children and Youth Services Review 47: 86-91.

Shroff R (2017) Predictive Analytics for City Agencies: Lessons from Children's Services. Big data 5: 189-196. 10.1089/big.2016.0052

＜第二章＞

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, Tie-Yan Liu. "LightGBM: A Highly Efficient Gradient Boosting Decision Tree". Advances in Neural Information Processing Systems 30 (NIPS 2017), pp. 3149-3157.

Shi, Y., Ke, G., Shoukhavoun, D., Lamb, J., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ye, Q., Liu, T., & Titov, N. (2022). Lightgbm: Light Gradient Boosting Machine. R package version 3.3.2, URL https:// CRAN.R-project.org/package=lightgbm.

＜第三章＞

Ishida, M. & Kudo, T. (2022). RMeCab: interface to MeCab. R package version 1.10.

R Core Team(2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: https://www.R-project.org/.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. Journal of Statistical Software, 33, 1, 1-22. URL https://www.jstatsoft.org/v33/i01/.

Shi, Y., Ke, G., Shoukhavoun, D., Lamb, J., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ye, Q., Liu, T., & Titov, N. (2022). Lightgbm: Light Gradient Boosting Machine. R package version 3.3.2, URL https:// CRAN.R-project.org/package=lightgbm.

＜第四章＞

Scott(2018). SHAP documentation. © Copyright 2018, Scott Lundberg. Revision 45b85c18., URL:
https://shap.readthedocs.io/en/latest/example_notebooks/api_examples/plots/waterfall.html

Riccardo Guidotti. “Counterfactual explanations and how to find them: literature review and benchmarking”. Data Mining and Knowledge Discovery (2022).

Ramaravind K. Mothilal, Amit Sharma, Chenhao Tan, Authors Info & Claims. “Explaining machine learning classifiers through diverse counterfactual explanations”. FAT* ’20: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency.

Rafael Poyiadzi, Kacper Sokol, Raul Santos-Rodriguez, Tijl De Bie, Peter Flach. “FACE: Feasible and Actionable Counterfactual Explanations”. AIES ’20: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society.

Kentaro Kanamori, Takuya Takagi, Ken Kobayashi, Yuichi Ike, Kento Uemura, Hiroki Arimura, “Ordered Counterfactual Explanation by Mixed-Integer Linear Optimization”. Transactions of the Japanese Society for Artificial Intelligence, 2021.

Mokkink et al.(2019). COSMIN Study Design checklist for Patient-reported outcome measurement instruments, version 2019. URL: https://www.cosmin.nl/wp-content/uploads/COSMIN-study-designing-checklist_final.pdf

Kyle M. Lang, E. Whitney G., Moore, Elizabeth M., Grandfield (2020). A novel item-allocation procedure for the three-form planned missing data design, MethodsX, 7. https://doi.org/10.1016/j.mex.2020.100941.

Little, Todd & Rhemtulla, Mijke. (2013). Planned Missing Data Designs for Developmental Researchers. Child Development Perspectives. 7. 10.1111/cdep.12043.

Marcel A.J. van Gerven, Babs G. Taal, Peter J.F. Lucas (2008). Dynamic Bayesian networks as prognostic models for clinical patient management, Journal of Biomedical Informatics, 41, 515-529. https://doi.org/10.1016/j.jbi.2008.01.006.

Bringmann LF, Vissers N, Wichers M, Geschwind N, Kuppens P, Peeters F, et al. (2013) A Network Approach to Psychopathology: New Insights into Clinical Longitudinal Data. PLoS ONE 8(4): e60188. https://doi.org/10.1371/journal.pone.0060188

V. A. Shterev, N. S. Metchkarski and K. A. Koparanov, ″Time Series Prediction with Neural Networks: a Review,″ 2022 57th International Scientific Conference on Information, Communication and Energy Systems and Technologies (ICEST), 2022, pp. 1-4, doi: 10.1109/ICEST55168.2022.9828735.

Bhanja, Samit. (2020). Deep Neural Network for Multivariate Time-Series Forecasting. Advances in Intelligent Systems and Computing. 1255. 267-277. 10.1007/978-981-15-7834-2_25.

Koslovsky MD, Hébert ET, Businelle MS, Vannucci M. A BAYESIAN TIME-VARYING EFFECT MODEL FOR BEHAVIORAL MHEALTH DATA. Ann Appl Stat. 2020 Dec;14(4):1878-1902. doi: 10.1214/20-aoas1402. Epub 2020 Dec 19. PMID: 35386276; PMCID: PMC8982957.

＜第五章＞

Design Council, Framework for Innovation: Design Council's evolved Double Diamond.

URL:https://www.designcouncil.org.uk/our-work/skills-learning/tools-frameworks/framework-for-innovation-design-councils-evolved-double-diamond/

内閣官房 情報通信技術(IT)総合戦略室, サービスデザイン実践ガイドブック(β 版), 2018 年 3 月. URL:https://cio.go.jp/sites/default/files/uploads/documents/guidebook_servicedesign.pdf

松本・江間, 2021; 東京大学未来ビジョン研究センター 技術ガバナンス研究ユニット AI ガバナンスプロジェクト, 2021 年 6 月. URL:https://ifi.u-tokyo.ac.jp/projects/ai-service-and-risk-coordination/

2021 年 6 月 リスクチェーンモデル (RCModel) ガイド Version1, 東京大学未来ビジョン研究センター 技術ガバナンス研究ユニット AI ガバナンスプロジェクト.
URL:https://ifi.u-tokyo.ac.jp/wp/wp-content/uploads/2021/07/RCM_210705.pdf

リスクチェーンモデル(RCModel)ケース検討事例：Case06 再犯可能性の検証 AI, 東京大学未来ビジョン研究センター 技術ガバナンス研究ユニット AI ガバナンスプロジェクト.URL:https://ifi.u-tokyo.ac.jp/wp/wp-content/uploads/2022/01/RCModel_Case06_Verification-of-Recidivism-Possibility-AI_JP.pdf

松本・江間(2021). AI サービスに係る「実現すべき価値・目的」と「リスクシナリオ」の類型化, The 35th Annual Conference of the Japanese Society for Artificial Intelligence, 2021. URL(Last Access 2022/07/15):
https://www.jstage.jst.go.jp/article/pjsai/JSAI2021/0/JSAI2021_1E3OS8a04/_pdf/-char/ja

東京大学未来ビジョン研究センター 技術ガバナンス研究ユニット AI ガバナンスプロジェクト(2021 年 6 月). リスクチェーンモデル(RCModel)ガイド Ver1.0, URL(Last Access 2022/07/15):
https://ifi.u-tokyo.ac.jp/wp/wp-content/uploads/2021/07/RCM_210705.pdf

国立研究開発法人産業技術総合研究所(2022). 機械学習品質マネジメント リファレンスガイド, 2022 年 7 月 12 日 18:00 作成, 2022 年 7 月 14 日公開版. URL:
https://www.digiarc.aist.go.jp/publication/aiqm/aiqm-referenceguide-v1.0-jp.pdf