特集
Special Feature

［AI 判断の根拠を説明する XAI を使いこなす］

2 産業利用における説明可能 AI の使いどころ

坂元哲平　安部裕之　（株）NTT データ

説明可能 AI の実現要件

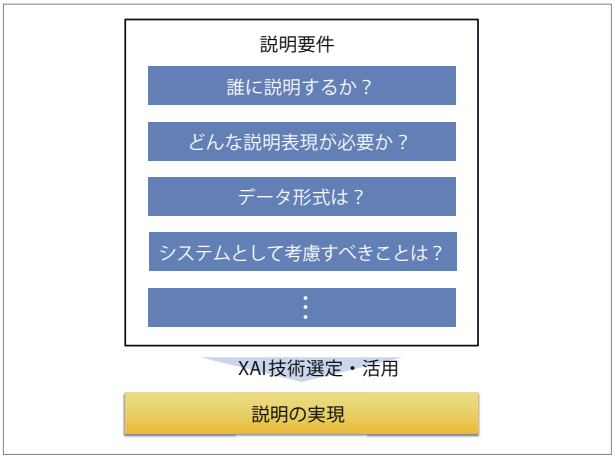
AI・機械学習技術の産業利用にあたって，AI の説明性や説明責任が 1 つの関心事となっており，説明可能 AI（eXplainable AI, XAI）技術が脚光を浴びている．

XAI の代表的な技術である LIME[☆1, 1)] は，2016 年に発表されている．その後，多くの XAI が提案され，2022 年現在では，技術についてはおおむね出揃ってきた状況にある．一方で，これらの技術によって AI の説明性の課題が解決されたとは言いがたいと筆者は考える．

XAI は「説明可能」という名称から，AI を説明できる魔法のような技術であると期待されることも多いが，現実は求められる要件と技術の間のギャップは小さくない．現在の XAI ができることを正しく理解するとともに，説明の要件を定め，要件に応じた技術選定や使い方の整理といった方法論の整備が必要である（図 -1）．

本稿では，個々の技術の説明は必要最低限にとどめ，説明要件に応じた XAI の使い方や留意点，課題について，データ形式，システム，そして人の観点から解説する．

☆1 ブラックボックスモデルの予測の根拠となった特徴量を提示する局所的説明技術を最初に提案した手法．説明対象データの近傍でのみ有効な（＝局所的な）ホワイトボックスモデルを構築することで説明を生成する．



■図 -1　説明の実現

AI の説明方式

この章では，今後の議論の準備として，AI の説明方式を整理する（表 -1）．なお，各技術の背景にある数理的な性質の深堀りはせず，説明表現（出力形式）をもとに整理を行う．

ホワイトボックスとブラックボックス

AI モデルには，説明力は強いが精度面で劣るホワイトボックスモデルか，説明力は弱いが精度面に利があるブラックボックスモデルの 2 つの分類がある．前者は，線形回帰モデルや，決定木モデル等が該当する．後者は，ランダムフォレストやブースティング等のアンサンブルモデルや，深層学習等が該当する．

ホワイトボックスモデルで十分な予測精度を達成するのであればそれでよいが，十分な予測精度を得られないことが多く，その場合，ブラックボックス

モデルを利用することとなる．ブラックボックスモデルの説明方式には大きく 2 つの分類があり，AI モデル自体を説明する大局的説明と，個々の予測結果を説明する局所的説明がある．

大局的説明

ブラックボックスモデル全体に注目し，全体傾向としてどの特徴量が重要かを計算したり，解釈可能なモデルで表現することにより説明する方法である．大局的説明の代表的な方式として，特徴量に注目する方式がある．その中でも，重要度を算出する方式は簡単でよく用いられる．Permutation Importance☆2 と呼ばれる技術によって AI モデルに依存せず算出できるほか，木構造モデルの Feature Importance☆3 のように，モデルの内部構造を用いて算出ができる場合もある．特徴量の重要度から一段踏み込んだものとして，特徴量の値と予測値の平均的な関係を表現する Partial Dependence☆4 も特徴量に注目した方式の 1 つである．

別の方式として，複雑なモデルをホワイトボックスモデルで置き換えた代理モデル（Surrogate Model）☆5 を作成することによって説明する方式もある．

局所的説明

特定の入力に対するブラックボックスモデルの予測の結果に対し，根拠を提示することで説明する方法である．個々の予測結果を説明する代表的な手法は，予測に寄与した特徴量を算出する方式であり，説明対象モデルに依存しない技術として LIME や SHAP☆6 が挙げられる．深層学習を用いたモデルでは，GradCAM☆7，Attention（注意機構）☆8 等も，計算方法こそ違うが，出力形式としては予測に寄与した特徴量を算出する技術と見なせる．

異なるアプローチとして，データに基づいた方式がある．この方式では，予測に寄与した学習データ，予測根拠が類似するデータ，予測結果が変わるデータといったように，データを用いて説明する．予測結果が変わるデータは，「もし○○なデータだったら」という意味から反実仮想データとも呼ばれる．

データ形式と XAI

本章では，データ形式と XAI について解説する．代表的なデータ形式として，表形式データ，テキストデータ，画像データを取り上げる．

☆2 サンプル間で特徴量の値をランダムにシャッフルし，予測精度がどれだけ低下するかを計測することで特徴量の重要度を計算する方法．

☆3 木構造のモデルにおいて，ノードの分割条件に使用した回数が多い特徴量や，良い分割ができる特徴量を集計することで，複雑なモデルを要約して説明する大局的説明技術．

☆4 特徴量の値を任意の範囲で変動させて繰り返し予測値を算出することで，特徴量の値と予測値の平均的な関係を表現したグラフを描画する大局的説明技術．

☆5 ブラックボックスモデルの出力を正解データとして，決定木等のホワイトボックスモデルを学習し，そのモデルをブラックボックスモデルの代理として説明する大局的説明技術．

☆6 ゲーム理論における個々のプレイヤの寄与度を算出する「シャープレイ値」がベースとなった，各特徴量の寄与の総和が予測値に一致する性質を持つ局所的説明技術．

☆7 畳み込みニューラルネットワークにおいて，入力を抽象化した情報である特徴マップと，出力ラベルに対する勾配を用いて，画像識別の判断理由としてハイライトを生成する局所的説明技術．

☆8 主にニューラルネットワークを用いた言語モデルにおいて，入力した系列の中で，遠く離れた要素からの影響を取り込むことによりモデルの性能を高める機構で，その影響度合いを局所的説明と見なすことができる．

■表 -1　説明可能 AI の整理

何を説明するか		どのように説明するか		技術例
ホワイトボックスモデル		（モデルそのものが説明可能）		線形回帰モデル，決定木モデル，k 近傍法
ブラックボックスモデル	AI モデル（大局的説明）	特徴量で説明	重要視する特徴量を算出	Feature Importance，Permutation Importance
			特徴量の値と予測値の関係を算出	PDP，ICE，ALE
		代理モデルで説明	簡単なモデルやルールで表現	Tree Surrogate，deflag Tree
	個々の予測結果（局所的説明）	特徴量で説明	予測に寄与した特徴量を算出	LIME，SHAP，Anchors，Grad-Cam，Attention
		データで説明	予測結果が変わるデータの例示	Counterfactual Instances，DiCE
			予測に寄与した学習データの例示	influence，MMD-critic

表形式データの説明

表形式データは，一般的なリレーショナルデータベースや，CSV ファイル形式で表せるデータである．多くの場合，特徴量一つひとつが意味を持つため，大局的説明も局所的説明も適用可能である．注意すべきは，特徴量の設計である．特徴量として元々のデータに含まれていたカラムだけでなく，精度向上のために，特徴量同士の比をとった変数や交互作用項を追加する場合がある．このような加工は特徴量エンジニアリングと呼ばれ，予測精度向上には寄与するが，必ずしも人間にとって解釈性の高い特徴量が用いられない場合がある．このため，高い説明性が必要とされるユースケースでは，特徴量設計の段階から説明性を考慮する必要がある．また，この問題は，元データからの加工プロセスを含めて XAI の説明対象とすることで解決できる場合がある（図 -2）．AI に投入される特徴量単位で説明すべきか，業務上解釈が容易な項目単位で説明すべきかを見きわめて説明を出力することが望ましい．

また，スパースデータ（ゼロ要素が多い高次元データ）を取り扱う場合は，非ゼロ要素のみで説明したほうが有効となる場合が多い．たとえば購買履歴データから顧客を分類するようなケースにおいて，ある商品を購入していないことを理由にするのではなく，購入した商品だけで説明するほうが直感的である．計算コストの観点からも，多数を占めるゼロ要素について考慮すると計算コストが非常に高くなることにも注意が必要となる．

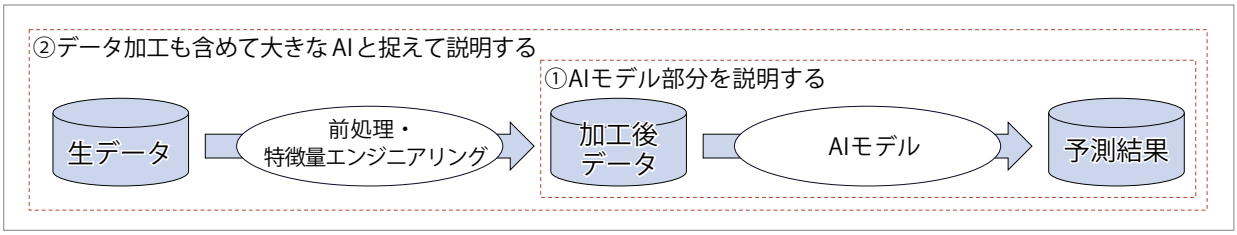
テキストデータの説明

テキストデータでは，基本的には単語が特徴量となる．単語そのものに意味があるため，大局的説明も局所的説明も適用可能である．局所的説明の際は，先述のスパースデータと同様に，出現した単語のみを説明対象とする．また，テキストデータでは文脈が重要な意味を持つため，説明の可視化の際にはもとのテキストデータにハイライトを重ねるようにすることが望ましい．

画像データの説明

画像データの場合，入力は各画素（画像上の座標）の RGB 値となる．このデータに対して大局的説明を考えたとき，AI モデル全体としてどこの画素が重要かを出力するが，1 枚 1 枚の画像によって対象物が写る位置は異なるため，画素単位の説明には意味がない．よって，各画像でどの領域に判断根拠があるかを示す局所的説明を用いる．局所的説明においても，画素単位では出力結果の解釈性が低く，計算量の観点でも問題があるため，画素をまとめた領域での説明が一般的である．まとめる際には，画像処理によって輪郭を抽出したスーパーピクセルを用いる方式や，単純に格子状に画像を分割する方式がある．また，LIME や SHAP 等のモデル非依存の手法は反復計算が必要なため，画像データに相性のよい深層学習では説明に時間を要する．そのため，深層学習の内部構造を活用した手法もよく用いられる．

他のデータ形式として，音声データやセンサデータもある．このような信号データは，時系列上の区間を特徴量としてハイライトする方式となる．また，



■図 -2　推論パイプラインと XAI

どのデータ形式においても，データで説明する方式は有効な手段となり得る．

AI システムから見た XAI

この章では，XAI の観点から AI システム構築における留意点を述べる．

1 点目は，XAI の AI モデルへの依存性である．たとえば，AI モデルにクラウドサービスの API を使用する場合には，モデルに依存する XAI の使用に制限がかかる．反対に，モデルに依存しない XAI を提供するサービスも近年あるが，この場合にはモデルの構造を利用した効率的な計算方法が使用できないという欠点もある．説明したいモデルと説明方法がシステムとして整合性が取れるかは検討事項となる．

次に，推論のパイプライン構築である．図 -2 の②のように，前処理や特徴量エンジニアリングを含めて AI モデルとして XAI に入力する際には，推論パイプラインとして扱いやすい設計になっている必要がある．XAI のために実装の手戻りがないよう，説明要件は早期に検討すべきとなる．

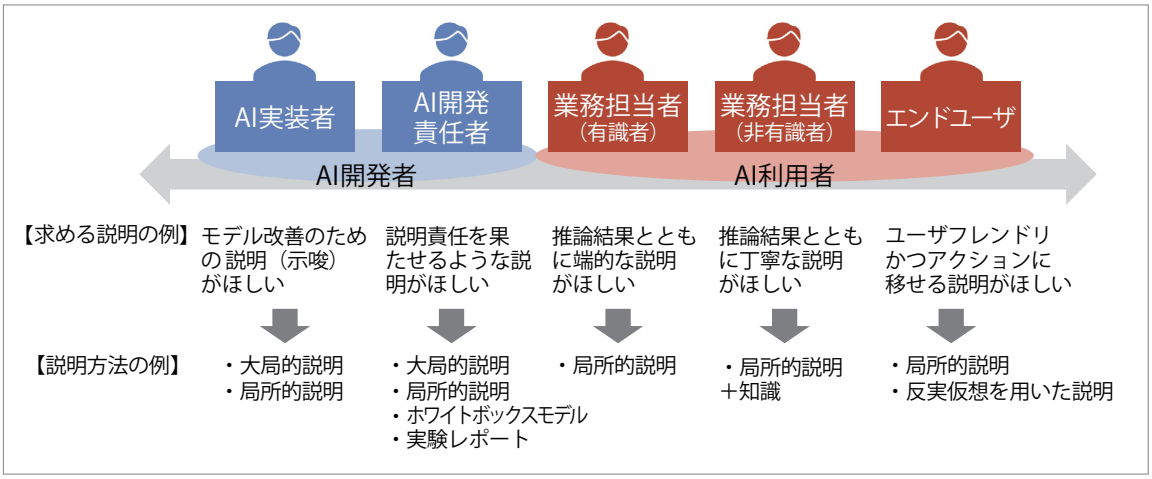
システムの応答時間にも考慮が必要となる．XAI には計算コストが高い方式もあるため，特にリアルタイム性が求められる場面では利用できる方式に制限がかかることを認識しておく必要がある．

最後に，画面・UI/UX である．AI に必要な情報を入力して，判定結果が出力されるシンプルなシステムでは，画面に関する考慮事項は比較的少ない．それに対し，XAI を必要とするシステムでは，画面の検討・開発に相応の工数を必要とする．納得感の醸成には，視覚的な情報が与える影響が大きいため，ユースケースに合わせて工夫を凝らす必要がある．

誰にとっての説明か

一般的な AI システムの周辺に登場する 5 つのペルソナ（典型的なユーザ像）を図 -3 に示す．ここで，それぞれのペルソナにとって求める説明内容が異なることに着目したい．それぞれのペルソナと説明方式がミスマッチしないように，誰に対する説明が必要かの要件定義をしたうえで，技術を選定すべきである．以降で，各ペルソナについて深堀りする．

AI の実装担当者は，AI モデル開発の中で，モデル改善を行う．モデル改善の際には，AI モデルの振る舞いを知ることが重要であり，XAI はそのヒントを得るために活用できる．たとえば，特徴量の重要度を算出し，業務知識と照らし合わせて違和感が



■図 -3　説明対象のペルソナと必要な説明

ないか，リーケージがないかなどを確認できる．もしも違和感があれば，データの取得方法や加工方法等を疑うことができる．また，開発段階で誤った予測をしたデータに対し，局所的説明によって誤り分析ができる．AI が苦手なデータの特徴を分析し，その結果をもとにデータ加工やモデリングを見直すことや，場合によっては例外処理で AI モデルには投入しないという判断を下すこともできる．

AI の開発責任者は，AI の構築方針の検討と，構築した AI のリリース判断を行う．前者において，きわめて高い説明性の要求がある場合には，ホワイトボックスモデルの使用も検討する．後者においては，判断のために実験・精度評価レポートが必要となり，その構成要素として大局的説明や局所的説明を利用できる．また，公平性の要求がある場合には，たとえば性別が予測値に影響を与えすぎていないかを確認することもできる．なお，AI 開発責任者にとっては，XAI による説明そのものが重要ということではなく，実験・精度評価レポート全体で，いかに納得感を醸成できるかが重要となる．そのため，XAI を用いずとも，多様なテストケースを設計し，各ケースで十分に精度を確認することでも，説明責任を達成できる場合もある．すなわち，XAI による説明に縛られず，リリース判断に足る説明責任とは何かを検討すべきであり，AI 品質管理技術も積極的に援用すべきである．

AI を利用する業務担当者（有識者）は，たとえば，融資審査 AI を利用する融資担当者のようなペルソナである．このペルソナは，AI モデルの全体傾向を知りたいわけではなく，業務上での個々のデータに対する予測の根拠が知りたいと想定できるため，局所的説明方式を選定すべきである．業務担当者に十分な知識や経験があれば，端的な説明で十分である．局所的説明の中でも予測に寄与した特徴量を算出する方式によって，重要な項目がハイライトされるだけで，AI の予測の根拠を類推することができると期待できる．同様に，類似データに基づく方式も有効となる．

AI を利用する業務担当者（非有識者）は，融資審査 AI を利用する新規配属の融資担当者のようなペルソナである．このペルソナにとっては，重要な項目がハイライトされても，なぜその項目が重要なのかが理解できない場合がある．よって，別の説明方式やアプローチが必要となる．1 つは，過去の類似事例をいくつか提示する方式である．この方式によって，目の前の事例だけでは理解が難しい場合でも，ある種の前例主義によって，理解・納得が進む可能性がある．別のアプローチとして，予測に寄与した特徴量に関する知識を提示することも有効である．融資審査の例では，ある財務指標がどのような意味を持つのかを提示し，その指標が高い（低い）と貸し倒れリスクが高まることが記載された文献を提示することで，理解をサポートできる．この知識は単純な検索技術でもよいし，知識グラフを用いてもよい．

最後に，AI を利用するエンドユーザ向けの説明について述べる．融資審査の例では，エンドユーザは借入申請をする顧客である．エンドユーザは自身に対する説明を求めるため，局所的説明技術を活用する必要があるだろう．ただし，提供する情報の粒度には留意が必要である．個々の特徴量の寄与度を提示するのではなく，特徴量を適切な単位でまとめ，可読性を増し，納得感を醸成する必要がある．また，予測結果を変えるためのヒントを得られるか，にも関心があるだろう．たとえば，融資審査の例で，特徴量に含まれていた「業種」が寄与して融資不可と判定されたとする．この判定結果と XAI の出力を顧客に伝えても，業種は簡単には変えられないため，受け入れられないだろう．この場合，制御可能な変数のみを用いて説明することが望ましい．たとえば，口座残高が少ないことが融資不可の理由であるとわかれば，顧客は残高を増やすための行動に移すことができる．さらに，反実仮想を用いた説明ができれば，具体的にいくらまで口座残高を増やせばよいか

が分かる有益な情報となる．また，XAI の出力をそのまま提示するのではなく，慣例的に説明してきたパターンに当てはめることも重要である．たとえば，あたかも担当者から説明されるように自然文で出力することで，顧客のシステムへの安心感を高めることができるだろう．簡易な方法として，テンプレート文に XAI 結果を合わせるだけでもよい．

以上，5 つのペルソナに必要とされる説明について解説した．ここでの解説はあくまで代表例であり，ユースケースによってはその限りではない．重要なことは，誰にとって納得できる説明かを定義した上で，XAI を活用することである．

XAI の限界と期待

本稿では AI に求められる説明性について，XAI をベースに，誰にとっての説明か，説明対象データはなにか，システム実装する際の考慮すべき観点を解説した．

現在の XAI は，AI の計算過程や予測の根拠を完全に説明するものではない．そもそも複雑な処理を簡潔かつ完全に説明することは不可能であることを前提にすべきであろう．また，人間にとって必ずしも理解が容易な出力が可能とも限らない．ある意味で無機質に提示される情報を解釈するのは人間の仕事となる．

最後に，XAI は魔法の技術ではないが，適切に使えば有用な技術であり，説明要件に整合した技術選定・技術開発が重要となる．そのための勘所・方法論の確立が期待される．

参考文献

1）Ribeiro, M. T., Singh, S. and Guestrin, C. : Why Should I Trust You? : Explaining the Predictions of Any Classifier, Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ADM, pp.1135-1144 (2016).

（2022 年 4 月 28 日受付）

■坂元哲平　Teppei.Sakamoto@nttdata.com

2018 年早稲田大学創造理工学研究科経営システム工学専攻修了．同年，NTT データ入社．AI 技術の社会実装に向けた研究開発に従事．共著に「XAI（説明可能な AI）そのとき人工知能はどう考えたのか」リックテレコム（2021）．

■安部裕之　Hiroyuki.Abe@nttdata.com

1998 年東北大学情報科学研究科システム情報科学研究科修了．同年，NTT データ入社．システムグランドデザイン，パッケージビジネス企画，データ分析・活用プロジェクトを経て，2021 年より AI 品質管理にかかわる研究開発に従事．