株式会社日立製作所

複雑なブラックボックス型 AI を判断基準が明確な AI に変換する AI 単純化技術を開発

お客さまに寄り添った信頼できる AI の実現に貢献

図 1-1 AI 単純化技術の概要

図 1-2 想定アプリケーション

日立は、従来のブラックボックス型 AI を判断基準が明確な AI に変換する AI 単純化技術を開発しました。従来のブラックボックス型 AI は、予測精度を高めるために複雑な数式で構成されており判断基準が不明確であるため、未知のデータに対して意図しない予測結果を導く不安やリスクがありました。一方、本技術を用いた AI は、あらゆる入力に対して人が理解できる単純な予測式を創出することにより、明確な判断基準の下で予測結果を提示します。さらに、お客さまの経験や知識に基づき予測式を調整できるため、予測精度を維持・向上しながら安心して AI を利用することが可能となり、お客さまに寄り添った信頼できる AI[*1] の実現に貢献します。

本技術の一部は、日立グループにおける製品出荷前の自動検査ラインに適用され、熟練者不足の解消や検査速度の向上効果が確認されました。今後日立は、製造・金融・インフラ制

御などさまざまな領域で、信頼できるAIの実装とそれを通じた社会全体のデジタルトランスフォーメーション(以下、DX)加速に貢献していきます。

なお、本技術開発は日立のAI倫理原則*2の実践項目に定める透明性・説明責任重視に従った取り組みの一環です。

■背景および取り組んだ課題

近年、メタバース等に代表されるデジタル化の加速度的な進展、深刻化する地球環境の悪化や新型感染症の拡大など、企業を取り巻く環境は急速に変化してきています。このような中で、お客さまの迅速なDX実現を支援するために、日立は信頼できるAIの実装に向けた研究開発に取り組んできました。

これまで、深層学習等の技術の進歩によりAIの予測精度は向上しましたが、AIには精度だけでなく、説明性、透明性、品質、公平性など、信頼できるAIとして複数の要件が求められます。特に、予測精度を高めるため多くの変数や複雑な数式により構成されたブラックボックス型AIはお客さまがその内容を理解困難なため、AIを安心して業務に適用できないという説明性の問題が指摘されました。このような中、日立はAIの判断基準を多角的に分析し、人に分かりやすく説明するeXplainable AI(以下、XAI)の技術を開発し、お客さまのさまざまな業務に適用いただきその有用性を検証してきました*3。その結果、AIの予測精度が高く、その判断基準をお客さまに説明できた場合でも、限られた条件下でお客さまが想定できない予測結果をAIが出力する場合、お客さまはAIを信頼・保証できず、対策に多くの時間を取られてしまうことが分かりました。

そこで日立は、これまでさまざまな領域のお客さまのDX支援で培ってきた知見を活かし、XAIの新たな方式として、複雑なブラックボックス型AIを判断基準が明確なAIに変換するAI単純化技術を開発しました。

■発表する論文、学会、イベントなど

本成果の一部は、2021年12月4日～7日にオンラインで開催されたIEEE SSCI 2021で発表されました。

■開発した技術の詳細

1. ブラックボックス型AI変換技術

深層学習モデルや勾配ブースティング木*4など、予測精度が高い一方で内容が複雑なブラックボックス型AIを、判断基準が明確なAI(単純な予測式)に変換する技術を開発しました。まず、従来のXAI分析技術により、AIへのさまざまな入力データに対し、要因(特徴量)が予測値に及ぼす影響の強さ(貢献度)を算出します。次に、特徴量が変化しても予測値への貢献度が一定である入力データの領域をクラスタリング技術により抽出します。抽出された入力データ領域ではその特徴量は予測値に影響を与えないため、その領域で判断基準を単純化で

きることを期待して AI を単純な予測式に変換します。このような処理を全ての入力データ領域で繰り返し、全領域をお客さまが理解可能な単純な予測式に変換します。

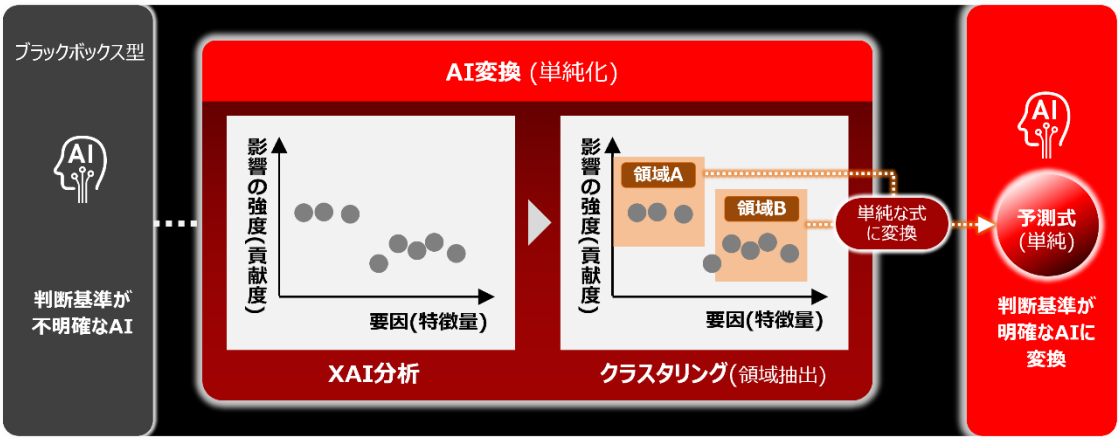

図 2 ブラックボックス型 AI 変換技術

2. 信頼性向上のための単純さ調整技術

必要な予測精度が得られるように、お客さまの経験や知識に基づき AI と対話・協調しながら、お客さま自身で予測式を調整する技術を開発しました。上記の技術1で得られた単純な予測式に対して本技術を用いることで境界値や各領域における予測式を調整することができます。

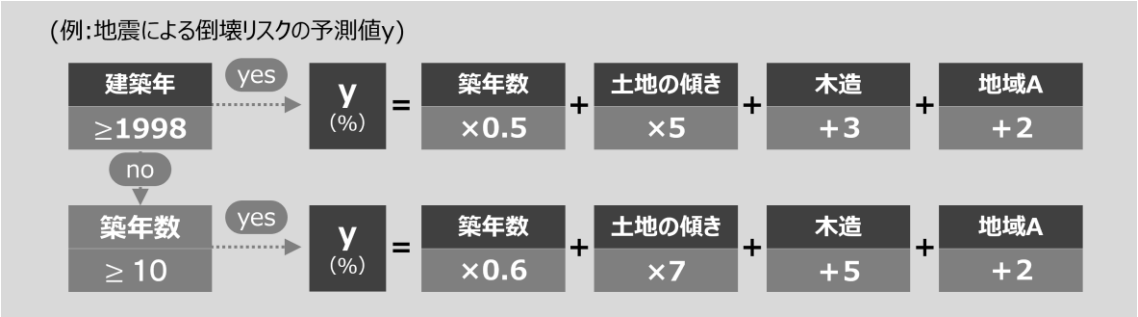

図 3 判断基準が明確な AI（調整前）の例

A. 上記の技術 1 で説明した AI の入力データ領域を分割する際の境界値を、お客さまの知識に合わせて調整可能です。例えば、図 3-A に示すように「耐震基準が 2000 年に変更されたので、1998 年ではなく 2000 年の前後で判断基準が替わるようにしたい」という調整が可能です。このように境界値をお客さまの知識と整合させることで、データが少ない領域での予測精度を向上させることができます。

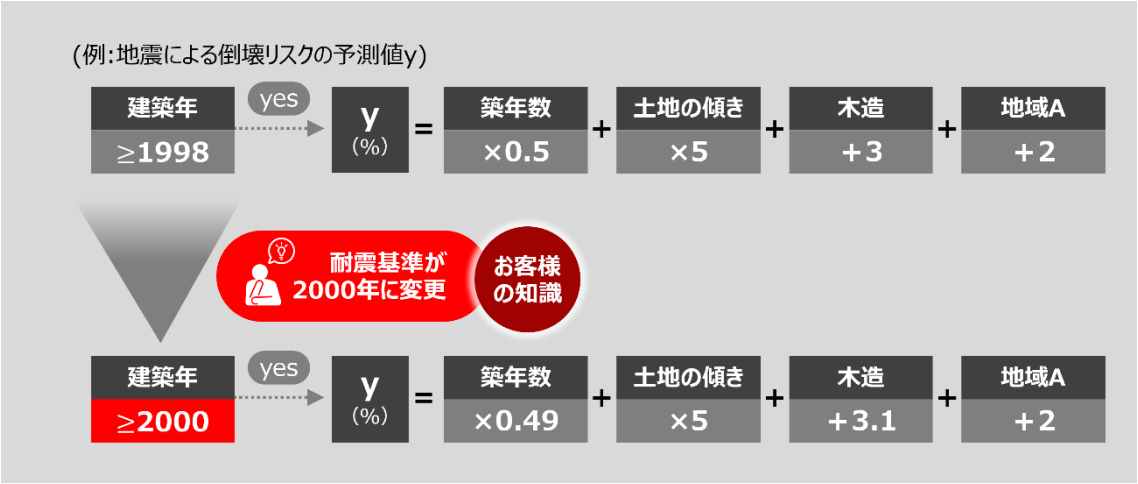

図 3-A 調整方法 1（お客さまの知識で境界値を調整）

B. お客さまは、各領域における予測式として、業務内容に合わせた式の形を指定可能です。お客さまが予測精度より予測式の単純さを優先する場合は、より単純な形の式を指定できます。図 3-B に示す例のように「予測式:y=築年数×0.5+土地の傾き×5+「木造なら+3」+「地域 A なら+2」+…が複雑であるため、変数を 3 つ以下にしたい」というような指定が可能です。

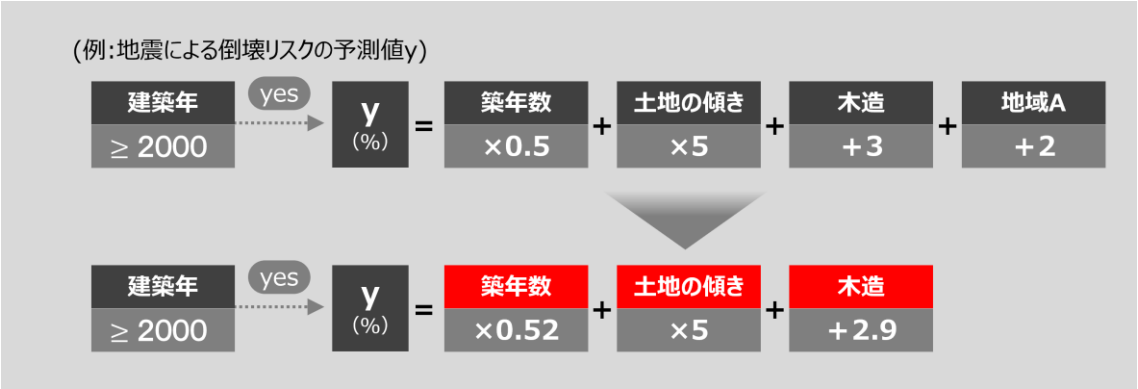

図 3-B 調整方法 2（関数系を指定することで単純さを調整）

上記 1 の技術で得られる式が単純だからこそ、お客さまの経験や知識を予測式に容易に反映させることができます。このように、お客さま自身で予測式を調整でき、信頼できる AI を容易・迅速に構築できます。

*1 信頼できる AI: https://www.hitachi.co.jp/rd/sc/ai-research/tech/xai/index.html

*2 日立の AI 倫理原則: https://www.hitachi.co.jp/New/cnews/month/2021/02/0222.html

*3 https://www.hitachi.co.jp/New/cnews/month/2020/01/0127.html

*4 勾配ブースティング木: 勾配降下法、アンサンブル学習、決定木の 3 つの手法が組み合わされた機械学習の手法。

以上