ゴール指向要求分析とシステム安全分析を利用した AI システム品質の個別ガイドライン導出方法の提案

Individual Guideline Derivation Method in AI System Quality Assessment by use of Goal-Oriented Requirements Analysis and System Safety Analysis

研 究 員：相津 一寛（パナソニック株式会社）
小宮山 英明（コニカミノルタ株式会社）
柳原 靖司（ブラザー工業株式会社）

研究概要

本論文では，AI システムの品質を保証する手段として，IGDM-AIQA 法(Individual Guideline Derivation Method in AI system Quality Assessment)を提案する．現在，AI システムは多くの分野で開発，運用されているにも関わらず，その特性から品質保証の方法が確立されていない．このような問題に対し，AI 開発の知見を集約してガイドライン化することが議論されているが，これらガイドラインは有識者向けで抽象度の高い内容となっており，AI の知見が必ずしも十分でない品質保証担当者が効果的に活用できるとはいえない．IGDM-AIQA 法を用いることで，対象システムの要件に基づいて品質アセスメントに必要な観点を導出し，品質保証の現場担当者が精度良く品質アセスメントを行える．

1. はじめに

機械学習技術の著しい発展により様々な産業分野で AI システムが開発，利用され始めているが，AI システムが内包する特有の性質から従来型の品質保証アプローチが通用しない課題が議論されている．このような課題に対して，AI システムに特化した品質保証のあり方が研究され，様々な AI システムの品質保証ガイドラインが発行されている．しかし，機械学習技術の応用分野は幅広く，既存の品質保証ガイドラインでは AI システムに共通する上位水準の知見，又は個別産業で共通する知見を提示しているのが実際である．これらガイドラインの記述は抽象度が高く，機械学習技術に精通していない品質保証担当者では記述の解釈が難解である．

本研究では，ゴール指向要求分析手法のひとつである AGORA[1](Attributed Goal-Oriented Requirements Analysis method)と，システム安全分析手法のひとつである FRAM[2](Functional Resonance Analysis Method)を応用して，AI システムの要求から品質保証の点で確認すべき項目（サブガイドライン）を抽出する IGDM-AIQA 法を考案し，実用性を評価した．仮想のクレジットカード与信審査システム（以降，FinTech 与信判定システムと略す）を例とした第三者による実験では，機械学習技術に詳しくない技術者であってもサブガイドラインを活用して有効な指摘ができることが分かり，要求工学やシステム安全性向上の知見に基づいてサブガイドラインを導出する意義を確認した．

以下，本論文の構成を述べる．まず，2 章では現状分析と課題提起を行う．次に 3 章では関連技術について言及し，4 章，5 章で夫々，解決策の提案と評価結果を示す．6 章で評価結果に対する考察を行い，最後に 7 章で成果と将来への発展で結ぶ．

2. 解決すべき課題

2.1 現状分析

機械学習技術を応用したシステム開発では，外界から収集された学習データに基づきモデルを生成しながら開発を進めて行く帰納的手法を採用しているため，従来の演繹的手法

に基づくシステム開発及び品質保証アプローチを適用できない．このような状況に対応するため，開発・運用フェーズにおいて AI システムの品質作り込みや品質評価の指針を与えるものとして，汎用的な AI システム品質のためのガイドライン（以降，汎用ガイドラインと略す）が発行されている．

2.2 課題提起

既存の汎用ガイドラインは，機械学習技術に対する知識が一定以上ある開発者を想定しており，多岐にわたる AI システムの応用分野における共通事項をまとめているため内容が抽象的で，機械学習技術の知見が十分でない品質保証担当者が活用することが難しい．機械学習品質マネジメントガイドライン[3]でも，「利用者が具体的な応用に即して，記述内容を取捨選択・具体化して用いることを想定している」と記されている．そのため，品質保証担当者でも精度良く品質の妥当性確認を行うための具体化されたガイドライン（以降，サブガイドラインと略す）を，システム要求に基づいて導出する枠組みがあるとよい．

3. 関連技術の説明

我々の研究の技術的拠り所として用いている，AI システム品質に関する汎用ガイドライン，AGORA 及び FRAM について説明する．関連技術として AGORA と FRAM に着眼した理由であるが，一般に AI システムでは要件の間でトレードオフが発生することが多いため，システムの要求から要件を獲得する過程を俯瞰的に可視化しながら分析できる点で AGORA，AI システムの運用時にプロジェクト関係者だけでなくエンドユーザや社会（コミュニティ）といった多様なステークホルダが関連するため，複雑な機能連関構造を可視化しながら分析できる点で FRAM が夫々適していると考えた．

3.1 AI システム品質に関する汎用ガイドライン

機械学習技術を利用したシステムやプロダクトの品質保証に対する共通指針を与えるものとして，機械学習品質マネジメントガイドライン，AI プロダクト品質保証ガイドライン[4]が発行されている．企業・大学等の有識者が知見に基づいて取りまとめたものであり，製品やサービスの開発者，利用者らが参照することを想定して記載されている．

3.2 AGORA

ゴール指向要求分析手法のひとつで，AND-OR ツリーグラフに属性値を付与した上で，主要求から下流に向かって副要求を展開しながら要件を導出する．リーフの各要件について，ステークホルダの満足度行列を調べることで主要求に対するゴール適合度や意見の対立度を計算することができる．図 1 に AGORA のモデル例を示す．

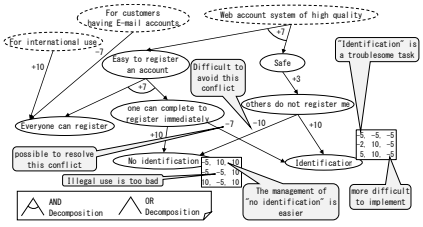


図 1 AGORA による要求分析例

3.3 FRAM

システム安全性向上におけるモデリング手法のひとつである．複数の機能がインタラクションする構造をモデルベースで分析することで，システムの長所や短所を特定できる．システムが正常に働くパターンを増強することを目的として利用する．図 2 に FRAM のモデル例を示す．

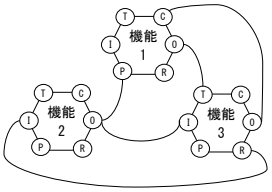


図 2 FRAM による分析例

4. 解決策の提案

4.1 課題の解決方針

汎用ガイドライン[3]では AI システムの品質観点を，ユーザに対する「利用時の品質」，システムの構成全体に求められる「外部品質」，構成要素の固有特性として定義される「内部品質」として捉えている．本研究では，当該ガイドラインにおける 3 つの品質観点を考

慮しながら，個別 AI システムに対する品質保証上の着眼点をビジネスゴールに適合するように導出する枠組み（IGDM-AIQA 法）を提案する（図 3）．また，現場の品質保証業務に対する IGDM-AIQA 法の有効性を評価するため，FinTech 与信判定システム（図 4）に適用した結果を示す．

【IGDM-AIQA 法の特徴】

・ゴール指向要求分析手法（AGORA）により利用時の品質と外部品質を考慮しながら AI システムを分析することで，該当システムに求められる要件群を目的指向で導出する．

・AI システムの有効性と公平性を定量的に解析する上で，与えられたデータセットに対する機械学習モデルのシミュレーションを行いながらデータ分析することで，当該システムの内部品質に求められる特性を把握する．

・AI システムの主要求に関わるステークホルダの機能連関構造（FinTech 与信判定システムでは公平性に配慮した社会受容性の構造）を明らかにする上で，各機能のインタラクションを FRAM により分析し（図 5），ステークホルダの重要度を特定する．

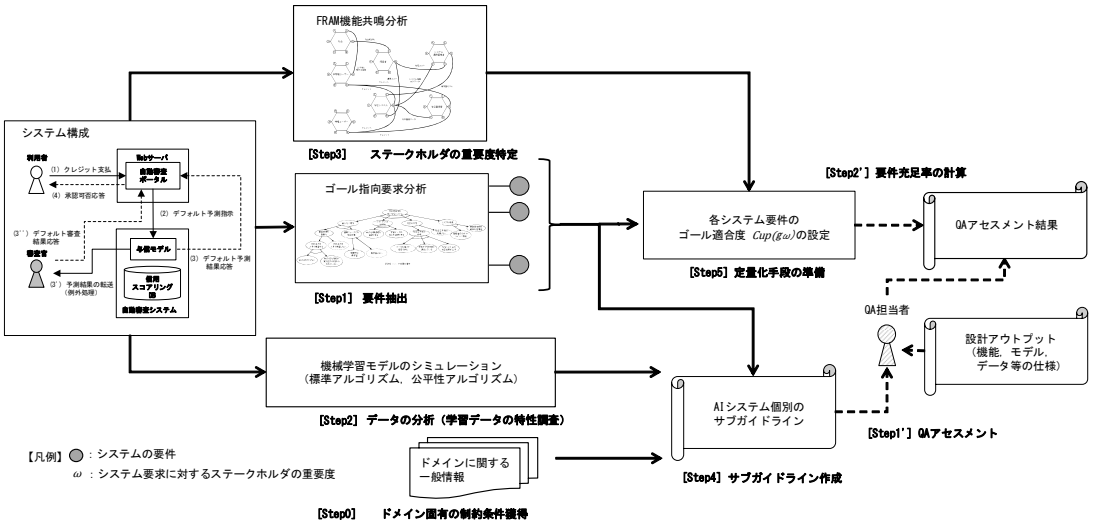


図 3 IGDM-AIQA 法（AI システム品質のサブガイドライン導出の枠組み）

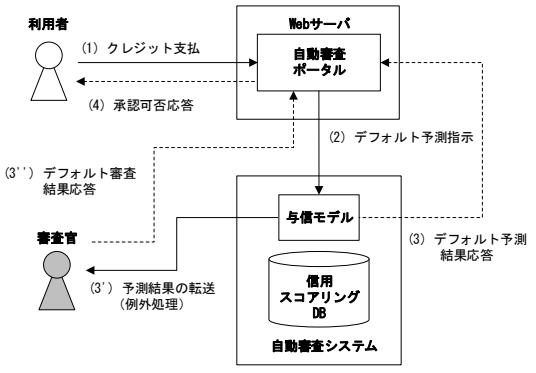


図 4 FinTech 与信判定システム

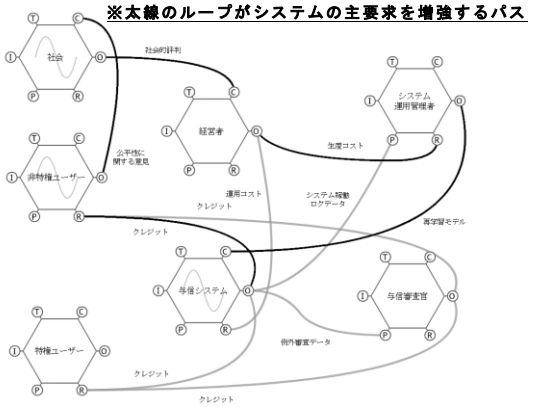


図 5 ステークホルダの機能連関分析(FRAM)

4.2 IGDM-AIQA 法を用いたサブガイドラインの導出手順

IGDM-AIQA 法は，AGORA，FRAM といったシステム開発で利用される既知の分析手法を活用しながら個別 AI システムに合わせて汎用ガイドラインを解釈するための枠組みである．品質保証業務に携わる QA 担当者は，表 1 に示す手続きに則って進めることで実用的なサブガイドラインを導出することができる．

表 1 IGDM-AIQA 法を用いたサブガイドラインの導出手順

手順	目的	□入力情報，■出力情報	手続き
STEP0	AIシステムの開発に付帯する制約の明確化	□システム仕様・設計情報，ドメインの関連資料[1] ■ドメイン固有の制約情報	ドメイン固有の情報から，目的システムの開発や運用に関わる制約情報（機械学習，運用，規制等の知見）を獲得する．
STEP1	AIシステムの品質観点を定義する際に参照する要件群の導出	□STEP0 の出力，汎用ガイドライン ■目的システムの要件群（品質観点の定義に適した粒度）	ゴール指向要求分析手法 AGORA を用いて，目的システムの主要求（ゴール）をツリー状にサブゴールへと展開しながら要件群を導出する．品質観点として，汎用ガイドラインに掲載されている内容（リスク回避性，AI パフォーマンス，公平性[2]等）を考慮する．
STEP2	機械学習コンポーネントの学習データに対する推論特性の調査	□目的システムが前提とする機械学習のデータセット ■機械学習アルゴリズムの推論特性，データセットの被覆性・均一性に関する情報	目的システムに搭載されている機械学習コンポーネントの学習データに対する推論特性をシミュレータ等で解きながら把握する．（機械学習の汎用アルゴリズムに対して Colaboratory(Google)，公平性アルゴリズムに対して AI Fairness 360(IBM)[8]，Fairlearn(Microsoft)[9]等を利用する）
STEP3	AIシステムのゴールに強い影響を及ぼすステークホルダの特定	□STEP0, 1, 2 の出力 ■Primary Stakeholder のリストとゴール（主要求）の貢献度に応じた重みω	システムの運用時に関わるステークホルダのうち，目的システムのゴール適合度に影響する Primary Stakeholder を FRAM 分析[3]により特定する．ステークホルダの機能連関構造を参考にしながら，システムの主要求に対する貢献度に応じて各 Primary Stakeholder の重みωを決める．
STEP4	QA アセスメントで使用するサブガイドラインの導出	□STEP1, 2, 3 の出力 ■サブガイドライン	システムの運用時に目的システムの機能が正しく発現されるための品質観点を，要件毎に汎用ガイドライン及び学習データの推論特性に基づいて検討し，ステークホルダの重要度を加味した上でサブガイドラインを導出する．
STEP5	AIシステムのゴール適合度$Cup(g_\omega)$の計算	□STEP1, 3 の出力 ■ゴール適合度	サブガイドラインを用いた QA アセスメントタスクの達成度を計測するため，図 6 に示す手順に基づき，目的システムのゴール適合度$Cup(g_\omega)$（式 1）を要件毎に求める．

[ゴール適合度の算出手順]

ゴール指向要求分析手法によって導出された要件のゴール適合度（貢献度）を定量化する手法が示されている[10]．本研究では，AI システムに関わるステークホルダ Stakeholder の内，主要な Primary stakeholder（UU, OW, OM）のゴール適合度を式(1)に基づき計算する．

$$Cup(g_\omega) \stackrel{\text{def}}{=} \frac{\sum_{s\in Stakeholder, p\in Primary\ stakeholder} \omega \cdot m(g)_{s,p}}{|Stakeholder| \cdot |Primary\ stakeholder|}. \quad (1)\ [10]$$

評価の視点（被評価者の立場）

要件名: 与信システムの判定結果が公平						
評価者（役割）	PU	UU	OW	OM	DV	役割毎の評価基準
PU	5	8	8	7	6	社会における機会均等性は理解できる
UU	8	10	9	8	6	より良い社会に向けて弱者に対する配慮が欲しい
OW	5	8	8	7	7	AI システムの性能と公平性はバランスが必要
OM	5	8	8	7	6	運用時に継続的にモデルを改善する必要がある
DV	4	7	7	6	6	技術によって不公平性を緩和することは難しい

STEP5-1: 満足度行列（左図）に各評価者の役割で，被評価者の視点に着目して-10～+10 の素点（要件重要度）を入力する．

STEP5-2: $Cup(g_\omega)$を計算する．[左図例 7]

分子: 左図のグレー部分の和　※役割毎に重み ω を付与（ω の値は STEP3 参照）[左図例 105]

分母: 左図の実線部と破線部に含まれる要素の集合濃度の積の平方根 [左図例 15]

[補足] 本研究では，研究員 3 名が夫々作成した満足度行列から$Cup(g_\omega)$を求め，これらを平均した．

【凡例】PU: Privileged User（特権ユーザ*1），UU: Unprivileged User（非特権ユーザ*1），OW: Owner（経営者），OM: Operations Manager（システム運用管理者），DV: Developer（開発者）

(*1) FinTech 与信判定システムでは PU が男性，UU が女性である．（女性の方が与信枠やデフォルト判定で不利）

図 6 $Cup(g_\omega)$導出のための満足度行列

4.3 仮説と研究設問

[仮説]

IGDM-AIQA 法から導出された AI システム品質評価のためのサブガイドラインを用いれ

[1] FinTech 与信判定システムの場合は，改正割賦販売法[5]，日本銀行のワークショップ報告書[6]，研究事例[7]等を参照した．

[2] 対象システムよっては公平性に対する要求が小さいため，適宜判断して除外する．（例：株価の予測，交通標識の識別）

[3] FinTech 与信判定システムの例では，図 6 の凡例に示すステークホルダのうち，図 5 の機能ループ（太線）を増強する重要なステークホルダとして，非特権ユーザ UU，経営者 OW，システム運用管理者 OM を選定した．なお，AGORA の$Cup(g)$の計算値に影響を与えるのは UU, OW, OM であるが，各ステークホルダの素点を同列で扱うのでなく，FRAM 機能連関構造内の「社会（六角形のオブジェクト）」に対する影響の仕方（強弱）を加味することにした．社会に直接作用する非特権ユーザ UU，社会からの出力を受けて指令を出す経営者 OW，間接的に機能ループの強化に作用するシステム運用管理者 OM という解釈を行い，3 名の研究員で合意して ω の重みを決定した．要件のゴール適合度を$Cup(g_\omega)$と再定義することで，ステークホルダの機能連関構造から生じる影響を要件の重みに反映させた（要件に対する重みは$\omega_{UU} = 1.0$，$\omega_{OW} = 0.9$，$\omega_{OM} = 0.8$を付与した）．

ば，ガイドラインがない場合，及び汎用ガイドラインを参照した場合に比べ，社会受容性を含むビジネスゴールを持った AI システムの品質保証のアセスメント精度が向上する．

[研究設問]

RQ1：機械学習技術に詳しくない技術者がサブガイドラインを参照すると，ガイドラインがない場合，及び汎用ガイドラインを参照した場合に比べ，システム要件に関わる欠陥指摘の精度が改善する．

RQ2[4]：機械学習技術に詳しい技術者がサブガイドラインを使っても，ガイドラインがない場合，及び汎用ガイドラインを参照した場合に比べ，システム要件に関わる欠陥指摘の精度は改善しない．

5. 解決策の評価

5.1 評価方法

仮説の検証は，FinTech 与信判定システムの概略設計資料（設計アウトプット）に含まれる欠陥事項をアセスメントするような枠組み（図 7）を用いた．独立した 2 つの被験者グループを用意することで RQ の定量的評価と，属性毎の定性的評価を行えるようにした． なお，5.2 節において前者は$\sum_i A_i \cdot Cup(g_{i\omega})$，後者は$\overline{A_i}$及び$\sum_i A_i$に着目して被験者のアセスメント精度を分析した．（$A_i$は要件 i に対する Assessment accuracy を意味する）

■ 実験意図の説明	⇒	■ 検証対象の提示	⇒	■ 実験の試行	⇒	■ 実験結果の評価
被験者に対し，システム要求（主要求と副要求）及びシステム構成図を提示し，（本実験の）品質アセスメントで問われている期待結果を理解させる．		設計チームが作成した不完全なシステム基本設計情報（表 2）を提示する．当該情報には，モデル，データ，その他の各区分で，設計意図が詳細に記述されている．		被験者グループ*2 Ⅰ群とⅡ群は，ガイドライン なし，汎用ガイドライン参照，サブガイドライン参照の各条件で，基本設計情報に含まれる欠陥を指摘する． ※30 分間で最大 10 個		被験者による指摘事項をシステムの要件（表 3）と照合しながら 3 段階で採点する*3. （A_i 1 点：指摘と理由が適合，0.5 点：指摘はできているが理由が不明瞭，0 点：指摘は要件の範疇外）

(*2) 被験者グループ Ⅰ群：機械学習技術に詳しくない技術者，Ⅱ群：機械学習技術に詳しい技術者
(*3) 要件に対する適合性で採点する．

図 7 IGDM-AIQA 法の効果測定の枠組み

表 2 システム基本設計情報（一部抜粋）

区分	項目	説明
モデル	アルゴリズム	2項分類問題を解くためのアルゴリズムとして，表現力が高く，正答率を高めやすいDNN（Deep Neural Network）を選定した．モデル構築のハイパーパラメータとして，BATCH_SIZE：25, EPOCH：20とした．

表 3 要件と対応するサブガイドライン

主要求：省人化に寄与する性能だけでなく，公平性に配慮した社会受容性の高い与信システム				
#i	要件	副要求	$Cup(g_{i\omega})$	サブガイドライン（要約）
1	ML*4 の計算過程を解釈できる	アウトプットの透明性	5.4	モデルのアルゴリズムは，説明性の高いアルゴリズムを使用しているか．
2	ML の汎化性能が実社会の状況から外れていない	省人化に寄与	4.7	モデルのアルゴリズムに含まれる汎化のために採用している制約によって，少数の重要なデータが無視されていないか．
3	データの偏りが受容できる	社会公平性	7.7	学習データの内容の分布が，偏っていないか．
4	標本が予測対象と適合している	社会公平性	6.5	学習データにクレジットカードのデフォルト予測で取り扱うすべての審査対象者のデータが網羅されているか．
5	学習データの加工を説明できる	アウトプットの透明性	3.5	正例（デフォルト），負例（非デフォルト）の不均衡を解消するため，近接データの内挿を行って，データを増やしているか．
6	ML の誤りが少ない	省人化に寄与	4.0	機械学習の推論結果に関する正答率，F1 値，AUC が十分であるか．
7	与信システムの判定結果が公平	社会公平性	7.1	・学習データの目的変数の値が性別で偏っていないか． ・推論結果が不公平な結果になっていないか． ・機械学習のバイアスを補正する処理が実施されているか．
8	機動的なモデルの再学習	省人化に寄与	2.9	再学習の時間は，運用で許容できる時間以内であるか．
9	管轄省庁のガイドラインに準拠	リスクの低減	2.9	収入が低い世代の人に対してのクレジット額が高くなっていないか．
10	事故発生時の解析の容易性	リスクの低減	3.0	学習，検証データと，モデルの学習履歴が必要な時に，確認することができるか．

(*4) ML は Machine Learning（機械学習）の略

[4] 本研究の構想段階において，IGDM-AIQA 法から導出したサブガイドラインは，機械学習技術に詳しくないが演繹的手法に基づくシステム開発には精通している技術者に対してのみ有効であると仮定した．一方，機械学習技術に詳しい技術者に対してはサブガイドラインの有効性が低く，既存の汎用ガイドラインを参照することで品質の作り込み活動を行えるものと考えていた．

5.2 評価結果

異なる業種（製造，情報・通信，金融）に属する計 13 社の被験者（機械学習技術に詳しくない技術者 I群: 25 名，機械学習技術に詳しい技術者 II群: 13 名）に対して，仮想 FinTech 与信判定システムのシステム概略設計資料をアセスメントしてもらい，当該資料に含まれる欠陥事項を日本語で指摘してもらった．（回答集計率 95%）

[RQ に関する評価結果]

システムの要求を期待結果とした前記システム概略設計資料のゴール適合度（欠陥指摘精度）を検証するため，各条件（条件 a: ガイドラインなし，条件 b: 汎用ガイドライン参照，条件 c: サブガイドライン参照）における被験者の回答結果（自然言語）を$\sum_i A_i \cdot Cup(g_{i\omega})$で定量化した（表 4, 5）．4.3 節に示す RQ を検証する上で，被験者回答データが正規分布に従っていると仮定した上で，条件 a と条件 c，及び条件 b と条件 c の各母集団に有意な差が無いことを t 検定（有意水準 5%）によって検証した．t 検定の準備として，F 検定によりサンプルの等分散性を確認したところ，I群については等分散性が無く，II群については等分散性があったので，夫々Welch の t 検定，Student の t 検定を用いた．I群に関する t 検定の統計量は条件 a と条件 c が 3.01×10^{-8}（＜0.05），条件 b と条件 c が 6.76×10^{-6}（＜0.05）であり，両者とも有意な差があることが分かったので RQ1 の妥当性が確認できた．II群に関する t 検定の統計量は，条件 a と条件 c が 8.27×10^{-5}（＜0.05），条件 b と条件 c が 1.33×10^{-3}（＜0.05）であり，両者とも有意な差があることが分かったので RQ2 の妥当性が確認できなかった．即ち，保有する機械学習技術の知見に依らずサブガイドラインがあれば，個別 AI システムの設計アウトプットの欠陥指摘精度が向上するので，条件付きで仮説を検証できたといえる．

表 4 被験者回答精度（I群$\sum_i A_i \cdot Cup(g_{i\omega})$）

Ⅰ群	条件a	条件b	条件c
平均	9.0	13.2	26.1
分散	17.8	37.9	102.9
要件充足率	18.9%	27.6%	54.7%
サンプル数	25	25	23

表 5 被験者回答精度（II群$\sum_i A_i \cdot Cup(g_{i\omega})$）

Ⅱ群	条件a	条件b	条件c
平均	16.8	19.7	30.6
分散	32.3	26.4	73.9
要件充足率	35.2%	41.3%	64.2%
サンプル数	13	12	12

[定性的な評価結果]

設計資料に含まれる欠陥事項に対する被験者の回答傾向を把握する上で，回答結果（自然言語）の要件適合性（$\overline{A_i}$及び$\sum_i A_i$）で可視化した．

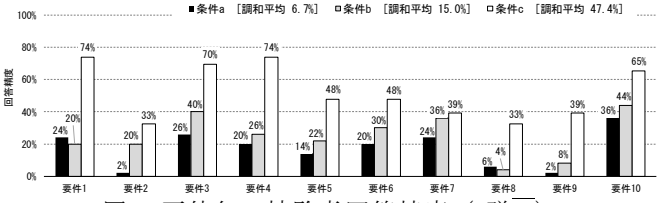


図 8 要件毎の被験者回答精度（I群$\overline{A_i}$）

<u>要件毎の被験者回答精度</u>

被験者回答精度を主要求への適合度という全体視点で見ると，サブガイドラインを参照した場合に FinTech 与信システムの基本設計情報に含まれる欠陥事項の回答指摘精度が有意に改善することが分かったが，要件毎に分解して検証すると改善幅に差が見られた（図 8, 9）．以下，被験者群毎に見られた特徴を示す．

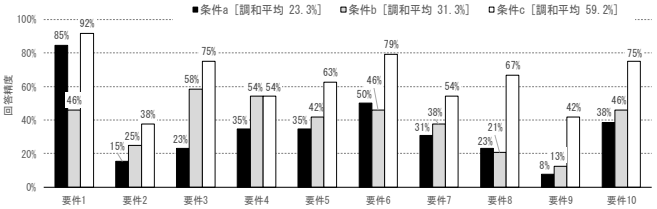


図 9 要件毎の被験者回答精度（II群$\overline{A_i}$）

Ⅰ群 $\overline{A_i}$	推定要因
改善幅が大きい要件：要件 1, 4（条件 a と c の差 50pt, 54pt）	サブガイドラインの対応箇所では機械学習アルゴリズムの種類，学習データの網羅性について言及しているが，機械学習技術に対する豊富な知識が無くても概念として理解しやすい．
改善幅が小さい要件：要件 7（条件 a と c の差 15pt）	サブガイドラインの該当箇所では機械学習の公平性について言及しているが，推論結果の偏りやそのバイアスを補正する考え方に対する理解が難しい．（特に後者は，研究領域として発展中）

Ⅱ群 $\overline{A_i}$	推定要因
改善幅が大きい要件：要件 3, 8（条件 a と c の差 52pt, 44pt）	機械学習技術に詳しい技術者は，機械学習アルゴリズムの種類と性能（要件 1, 6）を深く掘り下げて欠陥事項を指摘する傾向があり，学習データの偏りと機械学習の再学習時間（要件 3, 8）については関心が低かった．サブガイドラインの中で着目すべき視点を明示することで該当箇所にも意識を向けられる．

改善幅が小さい要件：要件 1（条件 a と c の差 7pt）	サブガイドラインの対応箇所では機械学習アルゴリズムの説明性について言及しているが，既に被験者が保有している機械学習技術の知見でも，基本設計情報に含まれる欠陥を容易に指摘できる．

共通	推定要因
条件 b（汎用ガイドライン参照）の精度が条件 a, c に比べて低い要件：要件 1, 8	要件 1:「説明性が高い機械学習アルゴリズム」という点について，個別システムに応じた解釈法が汎用ガイドラインでは解説されていないので精度が下がった． 要件 8:「機械学習の再学習」は当該分野の初歩の知識領域であるため，敢えてガイドラインには掲載されておらず，それ故，視点として着眼されにくかった．

品質保証におけるサブガイドラインの効果

所属企業での役割が QA 担当者（N=6）である被験者の回答結果を抽出して可視化したものを示す（図 10）．QA 担当者の場合，サブガイドラインを参照しながら欠陥指摘を行うと（条件 c），参照しない場合（条件 a, b）に比べ，夫々，4.2 倍（1.8pt→7.5pt），2.5 倍（3.0pt →7.5pt）の改善効果が認められた．この結果は，機械学習技術の有識者（II群）の対応する改善効果（1.6～1.7 倍）と比べても顕著である．

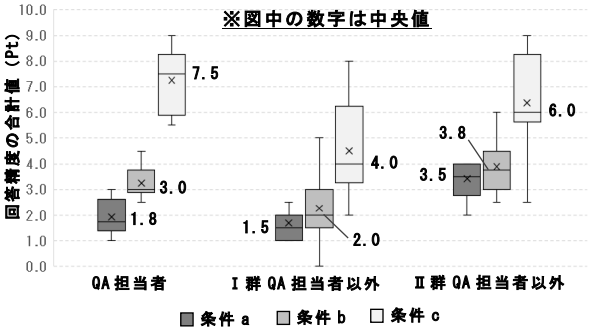


図 10 役割に着目した被験者回答精度$\sum_i A_i$

6. 考察

6.1 得られた知見

[仮説に対する整合性]（関連: 5.2 節 [RQ に関する評価結果]）

AI システムの品質保証活動において，基本設計情報の欠陥を要件群に照らして指摘するタスクについては，機械学習技術に関する知識や業務経験に依らず，要件群から導出した QA アセスメントのためのサブガイドラインに基づいて実施した方が精度を改善できる．記述の抽象度が高い，既存の汎用ガイドライン[3]を参照するよりも有意に改善する．

[サブガイドラインの記述]（関連: 5.2 節 要件毎の被験者回答精度）

個別 AI システムの要件群に適合するようなかたちでサブガイドラインを導出する際，機械学習技術に関する背景知識に依存して理解度にばらつきが生じないようにするため，IGDM-AIQA 法のサブガイドライン導出部で言語化の工夫が必要である．

[品質保証活動の現場での有効性]（関連: 5.2 節 品質保証におけるサブガイドラインの効果）

現場の実務でシステムやソフトウェアの品質保証活動を行っている担当者が，IGDM-AIQA 法から導出されたサブガイドラインを参照すると欠陥指摘の精度を 4 倍近く改善できるので，QA 担当者の本業の知見を補完しながら実務を遂行する上で合理性が高い．

6.2 IGDM-AIQA 法の実用性

[サブガイドライン導出の再現性]

IGDM-AIQA 法では機械学習，要求工学(AGORA)，及びシステム安全性向上(FRAM)の学際的知見を活用している．品質保証部門の現場で必要となる QA アセスメントタスクでは，各分野の緻密な理論を習得するよりは寧ろ欠陥検出やレビューに必要な視点を保有する方が重要であるため，初学者レベルの知識を保有すればよい．本研究のケーススタディに関わった 3 名の研究者は，計半年間の各分野の自習によりサブガイドラインの導出を行えた．なお，複数人の視点で抜け漏れを検証すれば，網羅性や十分性を担保できると考える．

[投資対効果]

品質保証の現場に IGDM-AIQA 法を展開する上で，従来の開発手法に慣れた QA 担当者に対して，前記 3 分野の初学者レベルの知識獲得を目的とした教育投資を行えばよい．これにより，ガイドラインを使わない場合(図 10 条件 a)の約 4 倍の QA アセスメント精度を見込める．他方，汎用ガイドラインを活用するためには，機械学習有識者の指導の下，プロジェクトでの活用を経験し，ガイドラインの咀嚼方法を学ぶ必要がある．経営的視点で機械学習有識者の資源を教育活動に振り向けることは，投資対効果の点で課題がある．

6.3 妥当性への脅威

・被験者の在籍している企業の多くは製造業や情報・通信に属しており，FinTech に特有の機械学習技術の知見は必ずしも持ち合わせていない．5.2 節の実験（条件 a, b）の試行において，ドメイン固有の知識を十分にもっていなかったことが不利に働き，条件 c の改善効果を見かけ上増長させた可能性がある．

・本研究では，FinTech（クレジットカードの与信審査）という，特定領域に対して IGDM-AIQA 法を適用したので，手法の汎用性を示す上ではさらなる実験が必要である．

7. まとめ

7.1 成果

本研究では，品質保証の現場の実務で活用しやすい AI システムの品質アセスメントのためのサブガイドラインを導出する枠組みとして IGDM-AIQA 法を提案した．FinTech 与信判定システムを事例に本手法から導出したサブガイドラインを品質保証ケーススタディに適用した結果，ガイドラインがない場合，及び汎用ガイドラインを参照した場合と比較してアセスメントの精度が向上することを確認した．特に，現場の QA 担当者がサブガイドラインを活用した場合に 4 倍程度の精度改善を見込めることが分かった．

7.2 将来への発展

本研究では，FinTech 与信判定システムを対象に IGDM-AIQA 法の有効性を評価したが，IGDM-AIQA 法によるサブガイドラインの導出方法は，特定のドメインに関係なく汎用的な手法であるため，他ドメインのシステムについても適用が期待される．

8. 謝辞

石川冬樹主査，栗田太郎副主査，徳本晋副主査には，多方面にわたり御指導を賜りました．また，研究コース５及び一般企業の有志の方に実験に御協力を頂きました．関係者の皆様に厚く御礼申し上げます．

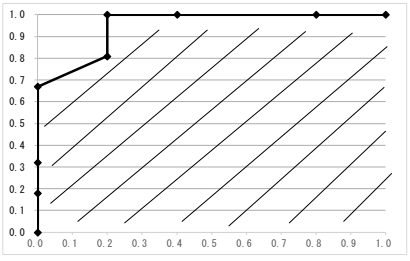
9. 参考文献

[1] H. Kaiya et al. (2002), AGORA: attributed goal-oriented requirements analysis method, 10th Anniversary IEEE Joint International Requirements Engineering Conference, pp.13-22.
[2] Erik Hollnagel, Örjan Goteman (2004), The Functional Resonance Accident Model, Cognitive System Engineering in Process Control 2004.
[3] 産業技術総合研究所, 機械学習品質マネジメントガイドライン第 1 版, https://www.cpsec.aist.go.jp/achievements/aiqm/ (閲覧 2020-12-27).
[4] AI プロダクト品質保証コンソーシアム, AI プロダクト品質保証ガイドライン 2020.08 版, http://www.qa4ai.jp/download/ (閲覧 2020-12-27).
[5] 経済産業省商務情報政策局, 割賦販売法, https://www.meti.go.jp/policy/economy /consumer/credit/11kappuhanbaihou.html (閲覧 2020-12-20).
[6] 日本銀行 金融機構局, AI を活用した金融の高度化に関するワークショップ 第 3 回, https://www.boj.or.jp/announcements/release_2019/rel190215d.htm/ (閲覧 2020-12-20).
[7] 小野潔 (2016)，インテックの与信モデルの特徴と今後の展開, ITJ2016.9 第 17 号.
[8] Rachel K. E. Bellamy, Kuntal Dey, Michael Hind et al. (2019), AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias, IBM Journal of Research and Development, Vol.63, Issue: 4/5, July-Sept. 2019.
[9] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík et al. (2018), A Reductions Approach to Fair Classification, In Proceedings of the 35th International. Conference on Machine Learning
[10] 佐藤慎一，石川冬樹，猪原健弘 (2011)，貢献度と顧客のニーズに関する妥当性の間のコンフリクト検出指標，ソフトウェアエンジニアリングシンポジウム 2011, pp.1-6.

付録 A　用語説明

※本論文で用いる用語の説明

表 A-1　用語説明

用語		説明
汎用ガイドライン		AI プロダクト品質保証コンソーシアム(QA4AI)，産業技術総合研究所（産総研）がリリースした，公になっているガイドライン．
サブガイドライン		本論文が提案している IGDM-AIQA 法で作成したガイドライン．
FinTech 与信判定システム		本論文のケーススタディとして仮想的に設定した，クレジットカード与信審査システム．クレジットカード利用者のデフォルト（債務不履行）を予測して承認可否を自動で審査する．
DNN (Deep Neural Network)		深層ニューラルネットワーク．パターン認識をするよう設計されたニューラルネットワークを多層構造化したアルゴリズムのこと．
BATCH_SIZE		上記 DNN で深層学習を行う際，損失関数を最小化するパラメータ（重み，バイアス）調整で入力のデータセットをいくつかのサブセットに分割する必要があり，このサブセットのデータ数のこと．
EPOCH		上記パラメータ調整では，損失関数が収束するまで学習を複数行うのが一般的で，この学習回数のこと． BATCH_SIZE，EPOCH の設定は，モデル精度に大きく影響を及ぼす．
ML (Machine Learning)		機械学習．コンピュータにデータを学習させ，特徴を発見して予測や識別をする．様々なアルゴリズムが考案されており，DNN は ML の一手法である．
F1 値		適合率と再現率の調和平均によって，モデルの性能を総合評価する指標．算出式は，以下のとおりである． F1 値＝(2*適合率*再現率)/(適合率+再現率)
	適合率	陽性と予測した内，実際に陽性であるものの割合．
	再現率	実際に陽性であるものの内，陽性であると予想した割合． 適合率と再現率はトレードオフの関係にある．
AUC (Area Under Curve)		ROC 曲線で，x 軸 y 軸で囲まれた部分（右図の斜線部）の面積の値．AUC が 1 に近いほど性能が高いモデルで，完全にランダムに予測される場合，ROC 曲線は原点(0,0)と(1,1)を結ぶ直線で AUC は 0.5 となる． 
	ROC 曲線	横軸に偽陽性率を，縦軸に真陽性率を置いてプロットしたもの．
F 検定		2 つのデータ群のばらつきが等しいか（等分散）を調べる方法．
t 検定		2 つのデータ群の平均の差が偶然誤差の範囲内にあるか否かを調べる方法．一般的に，有意水準 5%以内ならば有意差があると言える． データの正規性が確認され，F 検定の結果，等分散が仮定された場合に Student の t 検定を行う．データの正規性が確認され，不等分散が仮定された場合に Welch の t 検定を行う．
被験者回答精度		実験に参加いただいた被験者の指摘内容が，IGDM-AIQA 法で導出したサブガイドラインの要件にどこまで適合しているかの度合い．

付録 B　FinTech 与信判定システムに対する要求分析

※IGDM-AIQA 法によるサブガイドライン導出手順の補足（関連：論文 4 章）

１．要件抽出（STEP1）

ゴール指向要求分析手法（AGORA）を使った，要求分析の結果を示す.

主要求：社会受容性の高い与信システム

副要求：省人化に寄与，アウトプットの説明性，社会公平性，リスクの低減

要件：リーフの楕円（実線）が採用した要件群　※今回はケーススタディのため，10 個に限定

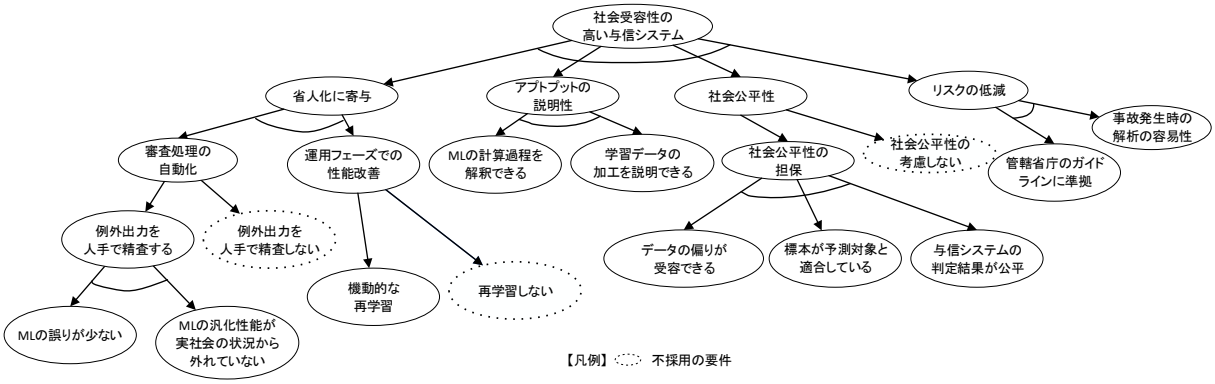


図 B-1　FinTech 与信判定システムに対する要求分析結果

２．ステークホルダの重要度特定（STEP3）

システムの要件を導出した後，運用フェーズにおいて各要件の実施（運用）に関わるステークホルダの機能連関構造を分析することで，ステークホルダの重要度を特定することができる．FinTech 与信判定システムでは，主要求を実現する上で副要求の「社会公平性」の貢献度が高いと仮定し，図 B-2 に示すような FRAM のモデリングを行った．具体的には，ステークホルダのひとつである非特権ユーザに着目し，非特権ユーザからの社会に向けた「公平性に関する意見」を継続的に改善することが，システムゴール（主要求）を増強すると考えた．同図の太線は，「社会からの評判に基づいて，経営者はシステムの運用改善に対するコスト負担を行い，システム運用管理者が機械学習モデルの公平性を増長するような再学習モデルを作成して与信システムに反映させることで，非特権ユーザの公平性に関する意見が良くなり，社会的評判に還元される」ような，ポジティブループのシナリオが表現されている．このシナリオでは，非特権ユーザ，経営者およびシステム運用管理者が重要なステークホルダである．

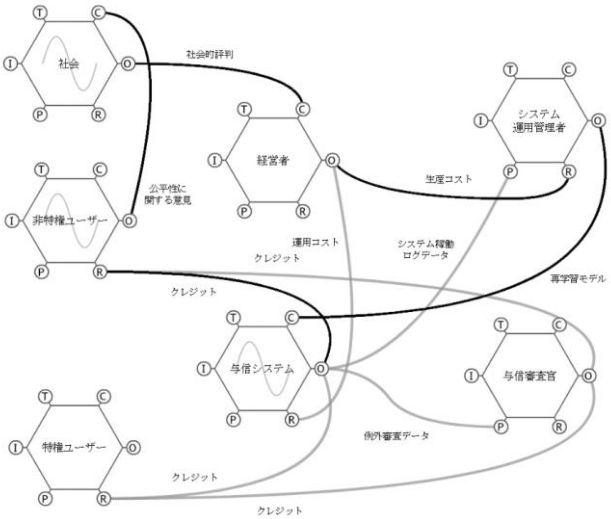


図 B-2　FinTech 与信判定システムに登場するステークホルダの機能連関構造分析結果

研究コース５（AI Quality Fairness チーム）

付録 C　実験概要

※被験者へ提示した資料（関連：論文 5 章）

１．はじめに

本研究では FinTech システムのケーススタディを通して，企業（ベンダー）の品質保証部門に属する担当者が現場で活用できる品質アセスメントのガイドライン作成手法を扱っています．既存の汎用ガイドライン（例：AI プロダクト品質保証ガイドライン，機械学習品質マネジメントガイドライン等）よりも，品質保証の対象となるシステム固有の特性を考慮しながら精緻にアセスメントできるように具体化している点が特徴です．（便宜上，我々が扱うガイドラインを「サブガイドライン」と呼びます）

２．実験の概要

被験者の方はベンダーの品質保証部門に属する QA 担当者になったつもりで，「FinTech 与信判定システム」の設計アウトプットの妥当性を評価（アセスメント）してください．このシステムでは，利用者によるクレジットカードの利用に際して，センター側にある自動審査システムが承認の可否をするものです．与信モデルにおける承認判断は機械学習アルゴリズムを使って自動化されていますが，承認判定（デフォルト予測）の信頼度が低い案件（トランザクション）については，センターに常駐する審査官が人手で判断します. システムの納品先となるユーザ企業の経営者からは，自社の業務プロセスを変革するにあたり，省人化による運用コストの削減に加え，昨今，世間を騒がしている AI の透明性や公平性の視点，与信システム固有の運用に関わる事故リスクの低減を要求されています．ベンダー企業とユーザ企業との間で合意形成した当該システムのゴールは「社会受容性の高い与信システム」です．

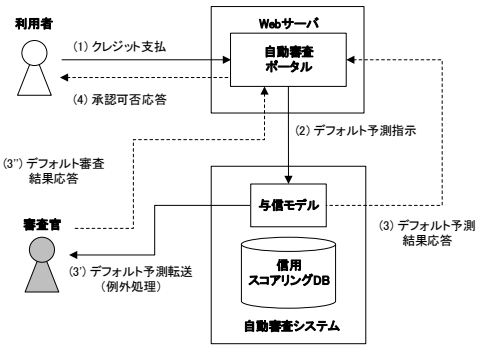


図 C-1　FinTech 与信判定システム

３．実験の進め方

＜全体像＞

各被験者には,インプット資料として FinTech 与信判定システムの簡易設計書を付録 D に提示します．また，被験者の QA アセスメントの検討資料として設計過程で試した様々な機械学習アルゴリズムの性能や，与信業務一般に関わるドメイン固有の情報をまとめた資料を付録 E に提示します．被験者は，これらのインプットをみながら気になる点とその理由を指摘してもらいます．（10 件を目安に指摘してください）

＜流れ＞

STEP1：ヒントなるガイドラインがない状態で QA アセスメントしてください．

[資料] 付録 D，付録 E

STEP2：産業技術総合研究所から公開された「機械学習品質マネジメントガイドライン[3]」の要約ドキュメント[1]を使って QA アセスメントしてください．

[資料] 付録 D，付録 E,「機械学習品質マネジメントガイドライン」の要約ドキュメント

STEP3：本研究で作成したサブガイドラインを使って QA アセスメントしてください．

[資料] 付録 D，付録 E，付録 F

※各 STEP，30 分以内で実行してください．指摘が 10 件に満たなくても構いません．

＜禁止事項＞

インターネットから，FinTech 与信判定システムに関する情報を知識として獲得しないでください．（機械学習アルゴリズムの一般的性質や金融与信業務の一般的要件については検索して頂いて OK です）

[1] 徳本晋(2020), 産総研による「機械学習品質マネジメントガイドライン」の調査, 非公開

４．被験者の属性情報の収集

本実験では，被験者を２つの群に分類して実験結果を集計する予定です．Ⅰ群は機械学習に詳しくない品質保証担当者，Ⅱ群は機械学習に詳しい技術者（研究者／開発者）です．適宜，属性情報と実験結果の関係を考察する可能性ありますので，あなた様の属性について教えてください．

4-a　所属する組織の概要（企業／大学，業種，役割）

4-b　機械学習の理解度（研究や実務での活用有無，経験年数，主な業務領域）

付録 D　FinTech 与信判定システムの概要

※被験者へ提示した資料（関連：論文 5 章）

概要：

このシステムでは，利用者によるクレジットカードの利用に際して，センター側にある自動審査システムが承認の可否を実行する．与信モデルの承認判断プロセスは，機械学習アルゴリズムを使って自動化されており，承認判定（デフォルト予測）の信頼度が低い案件（トランザクション）についてのみ，センターに常駐する審査官が人手で判断する．

要求分析：

システムの納品先（A 社）の経営者は，自社の与信業務プロセスの変革を実現する目的で既存の与信システムに FinTech 技術を導入することを決定したが，単に AI による省人化だけでなく，システムが社会的に受容されることを重視している．そこで，このシステムの要件を定義するにあたり，次の４つのサブ要求に着眼した．

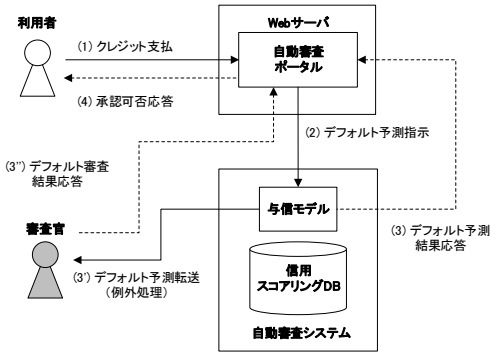


図 C-1　FinTech 与信判定システム（再掲）

１）省人化

従来，人手で実施していた与信のデフォルト予測業務を AI（機械学習）に置換することでコスト削減につなげられる．少なくともトランザクションの 8 割以上はコンピュータによる自動判定ができること．（判定の信頼度が低い 2 割のみ，従来のように人手で判定する）

２）アウトプットの説明性

与信判定業務に機械学習を導入することは省人化に寄与する一方，ステークホルダからのシステムに対する情報開示の要請に対して適切に対応する必要がある．具体的には機械学習の内部処理とアウトプットの関係性について合理的に説明できなければならない．

３）社会公平性

システムの運用において，技術面の優位性からもたらされるコスト削減効果だけでなく，近年，課題として取りあげられている AI 固有の倫理，道徳の側面にも配慮する．具体的には，年齢・性別・最終学歴等の信用スコアリングに影響するデータを機械学習で処理する際，社会的に許容可能な範囲で判定の公平性を担保する必要がある．

４）リスクの低減

管轄省庁の割賦販売法の規制によれば，過剰に与信を与えることで消費者の生活の営みに影響がでないように要求している．また，業界のガイドラインではシステム運用に際して事故を低減することが求められている．

システムの基本設計情報：

※表 D-1 のシステム基本設計情報には，意図的に欠陥を入れてある．

表 D-1　システムの基本設計情報（概略）

区分	項目	説明
モデル[2]	アルゴリズム	２項分類問題を解くためのアルゴリズムとして，表現力が高く，正答率を高めやすい DNN（Deep Neural Network）を選定した．モデル構築のハイパーパラメータとして，BATCH_SIZE：25, EPOCH：20 とした． ※モデルの性能については付録 E を参照のこと．

[2] アルゴリズムの性能，データ分析等の結果は，付録 D に掲載してある．

	前処理	学習データに対し，特殊な前処理は実行していない．
	後処理	判定結果に対し，特殊な後処理は実行していない．
	計算時間	対象データセットを用いたとき，初回のモデルの構築に要する計算時間は平均 14.960 秒であった．
	分類性能	未知データに対する分類性能は以下のとおり Accuracy（正答率）：0.817, F1 値：0.429, AUC：0.772
	汎化性能[3]	機械学習アルゴリズムの汎化性能が実社会の状況を適切に反映しているか否かは，「分類性能」の数値で妥当性を判断した．
	公平性指標[4]	性別に着目した判定結果の公平性指標（Equalized Odds Difference）：0.345
データ	選定	A 社のクレジットカード審査に通過し，実際に**クレジットカードを利用したユーザの過去のデフォルト状況**をデータセットとした．
	数量	データセットに含まれるデータ数は 30,000 件である．（内訳：正例 6,636 件，負例 23,364 件） ※デフォルトしたデータが正例である．
	分布	付録 E を参照のこと．
その他	運用	機械学習モデル更新は，オンライン学習方式を採用している．逐次，モデルを更新するため，**手元に学習データや学習履歴を保管しない．**

データセット（概略）：

FinTech 与信判定システムの「与信モデル（機械学習モデル）」の学習データとして，予測モデリングおよび分析手法関連の公開プラットフォームである Kaggle に投稿されてあるもの[5]を利用した．

表 D-2　学習データセット

サンプル数：30,000 件　説明変数：23 個　目的変数：1 個	
説明変数	各変数
クレジット	・LIMIT_BAL: Amount of given credit in NT dollars (includes individual and family/supplementary credit)
性別	・SEX: Gender (1=male, 2=female)
学歴	・ EDUCATION: (1=graduate school, 2=university, 3=high school, 4=others, 5=unknown, 6=unknown)
婚姻	・MARRIAGE: Marital status (1=married, 2=single, 3=others)
年齢	・AGE: Age in years
支払状況	・PAY_0: Repayment status in September, 2005 (-1=pay duly, 1=payment delay for one month, 2=payment delay for two months, … 8=payment delay for eight months, 9=payment delay for nine months and above) ・PAY_2: Repayment status in August, 2005 (scale same as above) ・PAY_3: Repayment status in July, 2005 (scale same as above) ・PAY_4: Repayment status in June, 2005 (scale same as above) ・PAY_5: Repayment status in May, 2005 (scale same as above)

[3] 未知のデータに対する識別能力のこと．汎化し過ぎるとデータの構造の大事な部分も無視してしまい，鈍感なモデルとなる．一方，敏感なモデル（表現力の高いモデル）ではデータの繊細な挙動まで学習してしまい，未知データの予測が難しくなる．

[4] Equalized Odds Difference=0 で判定結果が公平であるとみなす．

[5] https://www.kaggle.com/uciml/default-of-credit-card-clients-dataset (閲覧：2021-01-08)

	・PAY_6: Repayment status in April, 2005 (scale same as above)
請求額	・BILL_AMT1: Amount of bill statement in September, 2005 (NT dollar) ・BILL_AMT2: Amount of bill statement in August, 2005 (NT dollar) ・BILL_AMT3: Amount of bill statement in July, 2005 (NT dollar) ・BILL_AMT4: Amount of bill statement in June, 2005 (NT dollar) ・BILL_AMT5: Amount of bill statement in May, 2005 (NT dollar) ・BILL_AMT6: Amount of bill statement in April, 2005 (NT dollar)
過去の支払額	・PAY_AMT1: Amount of previous payment in September, 2005 (NT dollar) ・PAY_AMT2: Amount of previous payment in August, 2005 (NT dollar) ・PAY_AMT3: Amount of previous payment in July, 2005 (NT dollar) ・PAY_AMT4: Amount of previous payment in June, 2005 (NT dollar) ・PAY_AMT5: Amount of previous payment in May, 2005 (NT dollar) ・PAY_AMT6: Amount of previous payment in April, 2005 (NT dollar)
目的変数	各変数
デフォルト状況	・default.payment.next.month: Default payment (1=yes, 0=no) デフォルト(1): 6,636 件 非デフォルト(0): 23,364 件

付録 E　FinTech 与信判定システムが内包する与信モデルの学習データと性能

※被験者へ提示した資料（関連：論文 5 章）

（１）性別

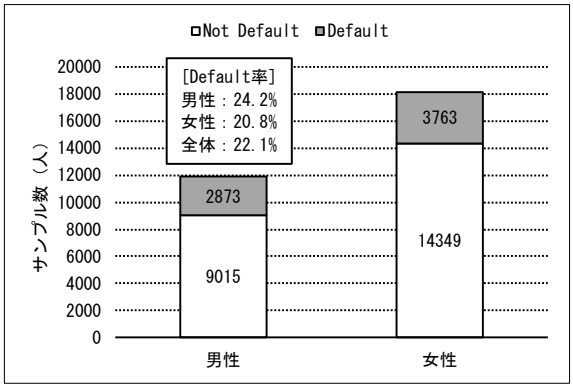


図 E-1　性別と Default（不履行）の関係

（２）最終学歴

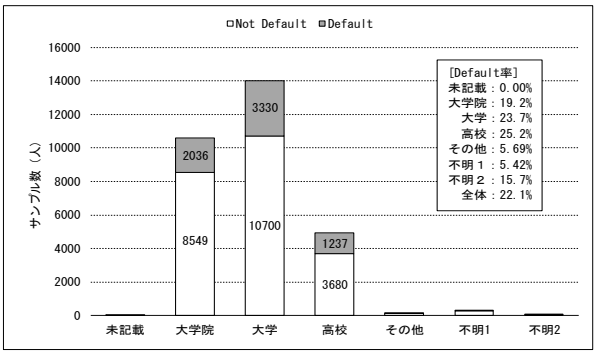


図 E-2　最終学歴と Default（不履行）の関係

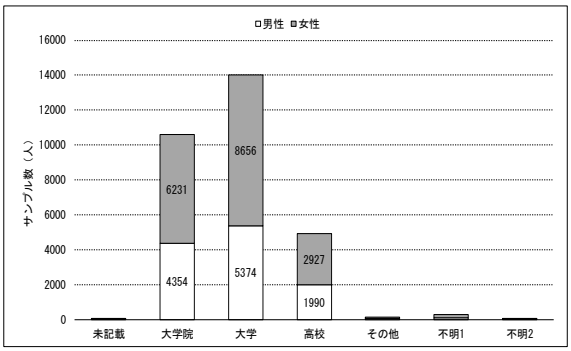


図 E-3　最終学歴と性別の関係

（３）婚姻状況

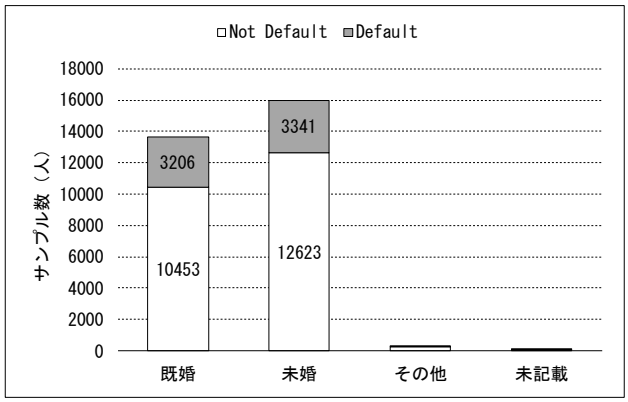


図 E-4　婚姻状況と Default（不履行）の関係

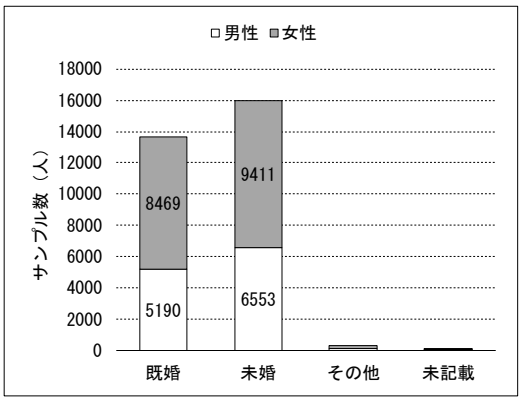


図 E-5　婚姻状況と性別の関係

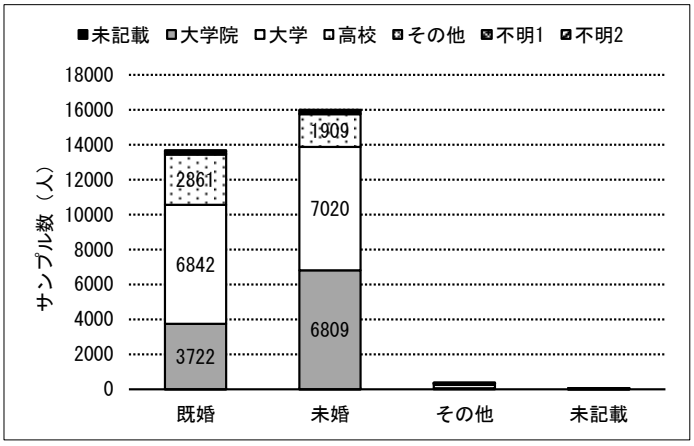


図 E-6　婚姻状況と最終学歴の関係

（４）年齢

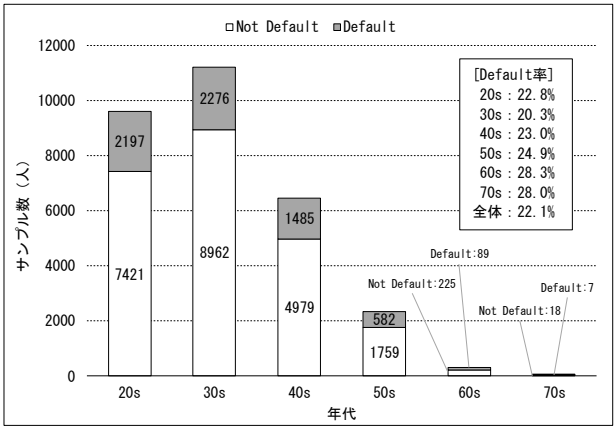


図 E-7　年代と Default（不履行）の関係

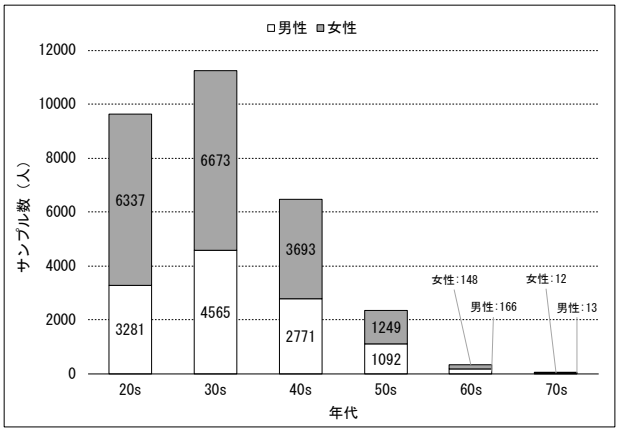


図 E-8　年代と性別の関係

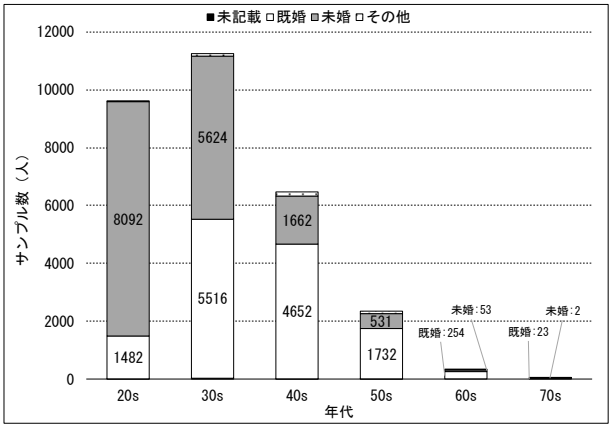


図 E-9　年代と婚姻状況の関係

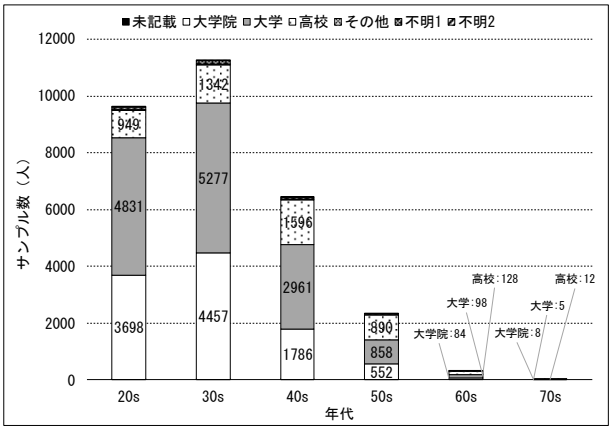


図 E-10　年代と最終学歴の関係

（５）与信額

※1 ニュー台湾ドル（TWD）= 3.64 円（2020 年 10 月）

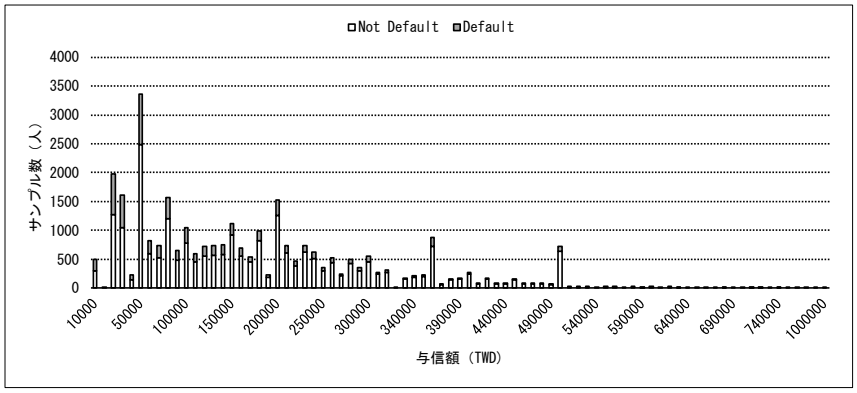


図 E-11　与信額と Default（不履行）の関係

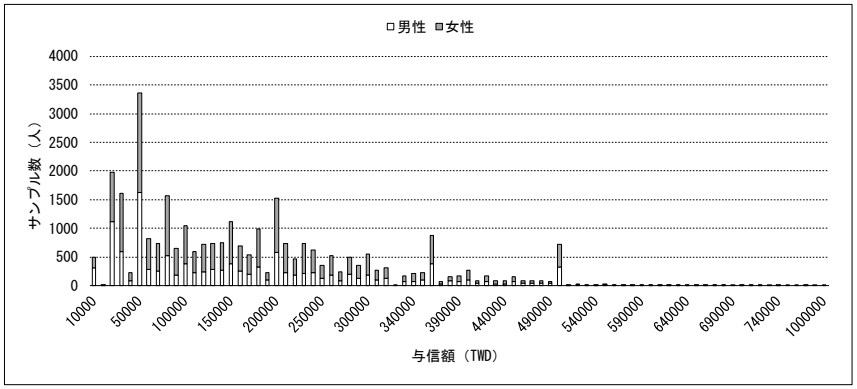


図 E-12　与信額と性別の関係

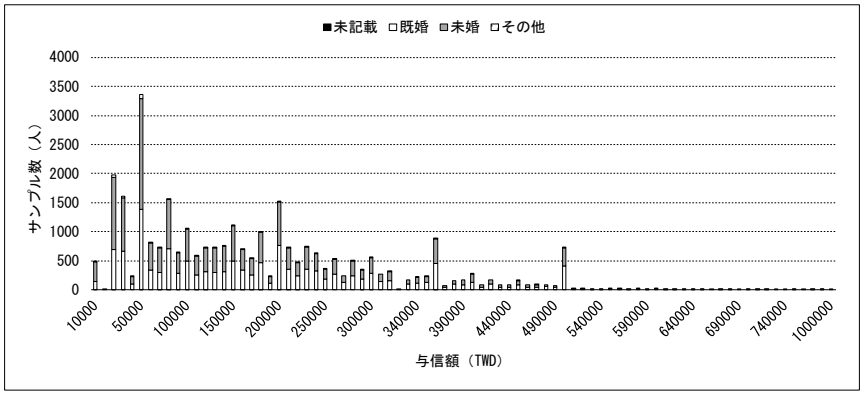


図 E-13　与信額と婚姻状況の関係

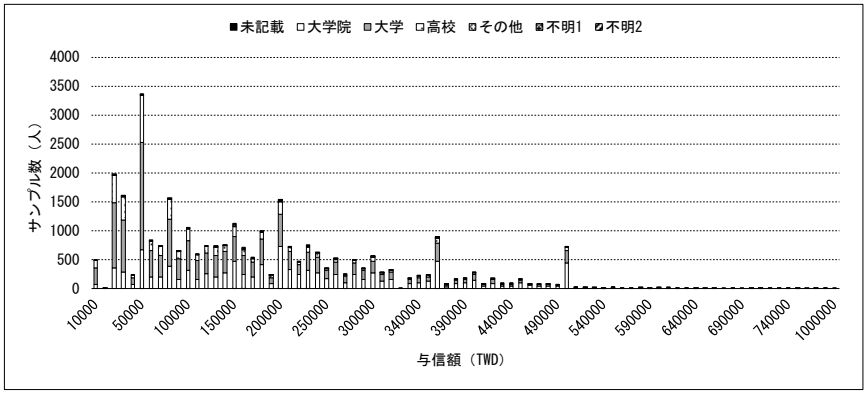


図 E-14　与信額と最終学歴の関係

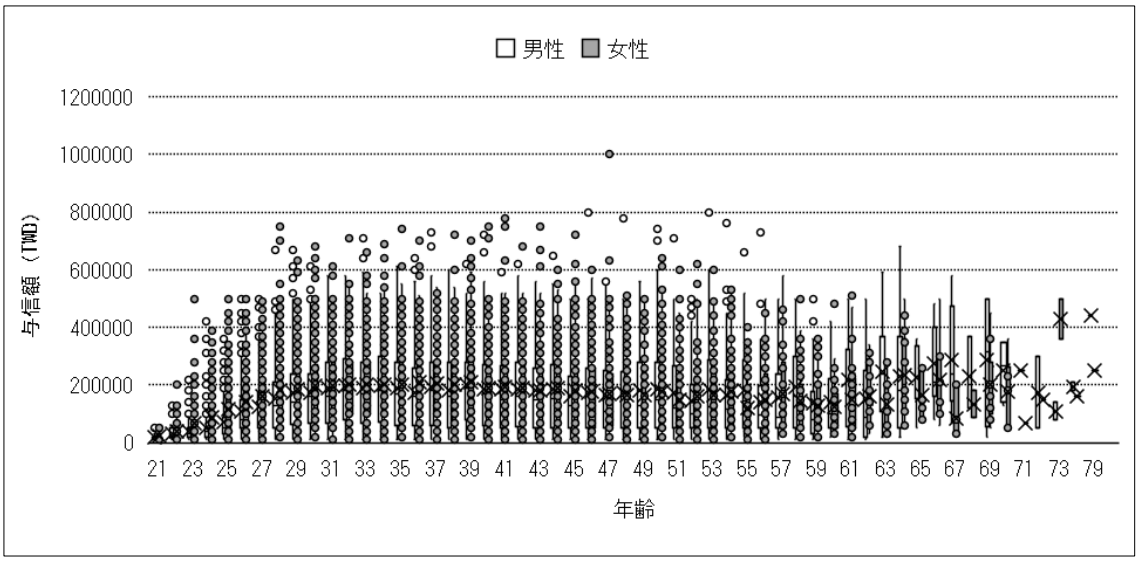


図 E-15　年齢，性別と与信額の関係（箱ひげ図）

（６）モデル性能

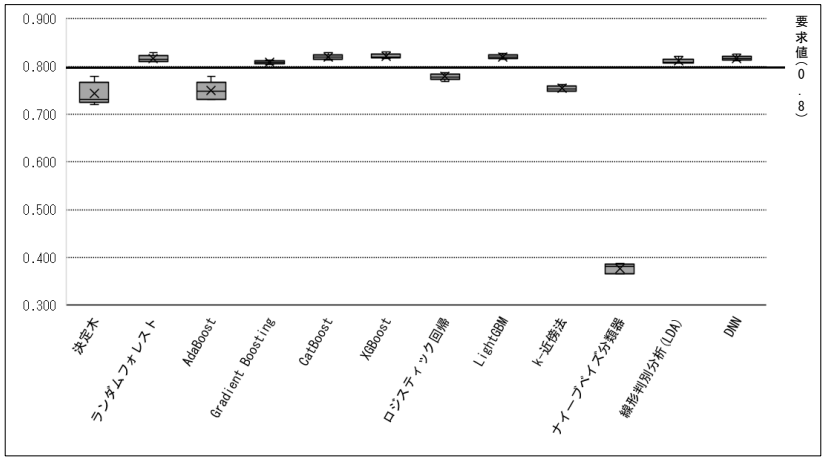


図 E-16　Accuracy（正答率）

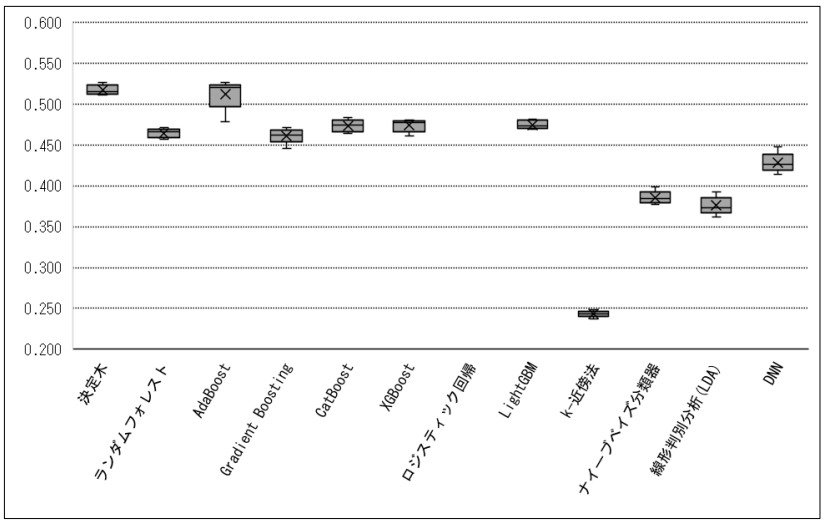


図 E-17　F1 値

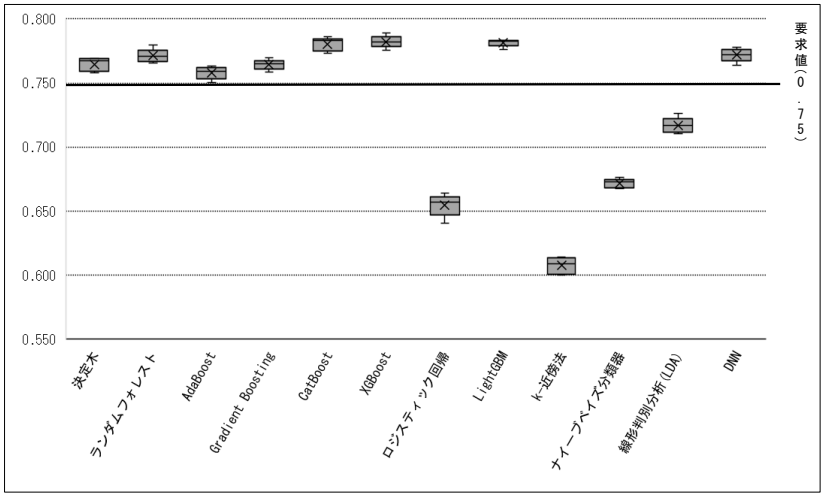
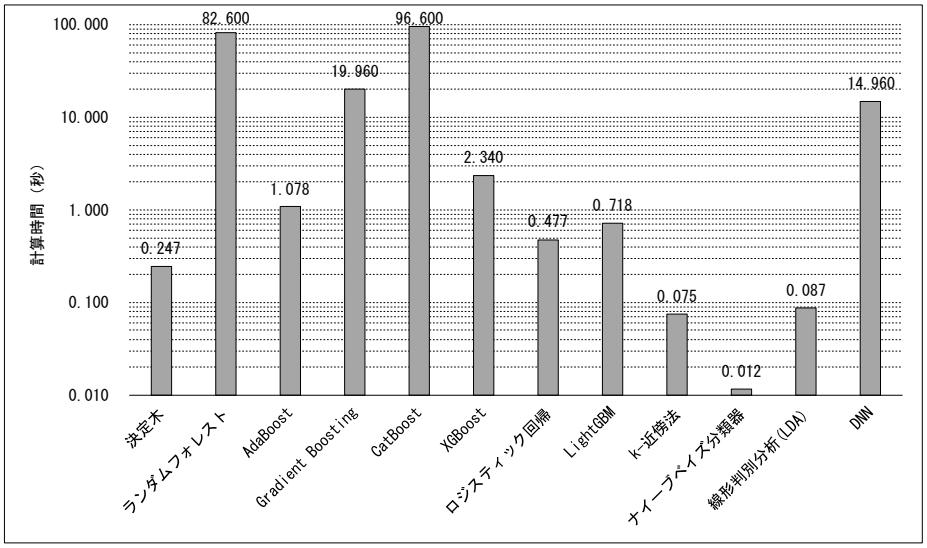


図 E-18　AUC



※Colaboratory(Google)のクラウドサーバにおける計算時間，交差検証（KFold=5）の平均値

図 E-19　計算時間

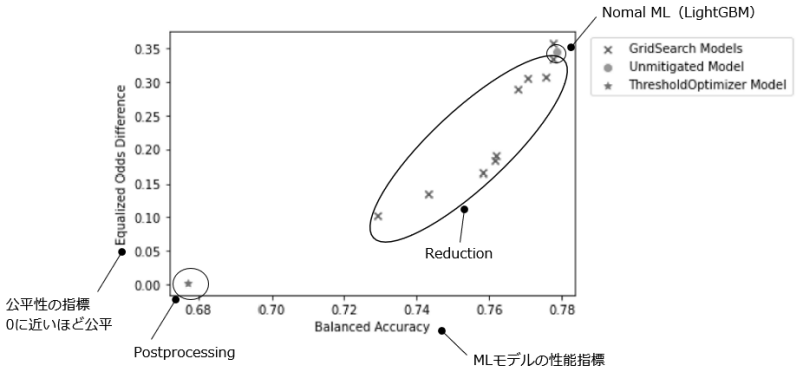
付録 F　FinTech 与信判定システムの品質保証のためのサブガイドライン

※被験者へ提示した資料（関連：論文 5 章）

FinTech 与信判定システムの品質保証活動（QA アセスメント）において参照すべき評価観点を具体的に記述してある．

表 F-1　サブガイドライン本体

#	項目	内容
1	要件	ML の計算過程を解釈できる
	確認事項	・モデルのアルゴリズムは，説明性の高いアルゴリズムを使用しているか．
	確認手段	・選択しているアルゴリズムは，デフォルトしやすいグループをツリー構造で表現し，分析結果が審査担当者にわかりやすいかを確認する．※一般的には決定木，XGBoost，LightGBM 等が該当する．（付録 D システムの基本設計情報を参照）
2	要件	ML の汎化性能が実社会の状況から外れていない
	確認事項	・モデルのアルゴリズムに含まれる汎化のために採用している制約によって，少数の重要なデータが無視されていないか．
	確認手段	・アルゴリズムの未知のデータに対して与信判定するために採用されている制約についての妥当性が検討されているかを確認する．（付録 D システムの基本設計情報を参照）
3	要件	データの偏りが受容できる
	確認事項	・学習データの内容の分布が，偏っていないか．
	確認手段	・学習データの分布図を確認して，特定のデータ値の偏りや統計データとの乖離がないかを確認する．（学習データの分布は付録 E を参照） －年齢データの比率（年齢ごとのデータ分布と，クレジットカード所有数の統計データと比較) －最終学歴データの比率（最終学歴の統計データと比較） －学習データ内のデフォルトの比率
4	要件	標本が予測対象と適合している
	確認事項	・学習データにクレジットカードのデフォルト予測で取り扱うすべての審査対象者のデータが網羅されているか．
	確認手段	・クレジットカードのデフォルト予測に必要なデータ種類（年収など）が学習データに含まれているかを確認する．（付録 D データセット（概略）を参照） ・20 代以上の年代の全データが含まれているかを確認する．（付録 E （４）年齢を参照） ・学習データセットの選定対象を確認する．クレジットカードの審査が通らなかった人のデータが含まれているかを確認する．（付録 D システムの基本設計情報を参照）
5	要件	学習データの加工を説明できる
	確認事項	・正例（デフォルト），負例（非デフォルト）の不均衡を解消するため，近接データの内挿を行って，データを増やしているか．
	確認手段	・学習データの前処理として SMOTE（Synthetic Minority Over-sampling TEchnique）等，正例と負例のデータ不均衡を解消するための処理を行っているかを確認する．（付録 D システムの基本設計情報を参照）
6	要件	ML の誤りが少ない
	確認事項	・機械学習の推論結果に関する正答率，F1 値，AUC が十分であるか
	確認手段	・未知データに対するモデルの分類性能を調べ，汎化性能が十分であるかを確認する．（付録 D（６）モデル性能を参照） －正答率をみて，経営者の要求を満足しているかを確認する． －F1 の値をみて，十分な値であるかを確認する．

		－AUC の値をみて，十分な値であるかを確認する．（一般的には，0.7 以上であれば，性能が高いとされている）
7	要件	与信システムの判定結果が公平
	確認事項	・学習データで正例（デフォルト）になっているデータは特定の性別に偏っていないか． ・システムが出力する判定結果が不公平な結果になっていないか． ・少数の非特権ユーザに対して不利な判定結果になるような，機械学習アルゴリズムのバイアスへの対処がされているか．
	確認手段	・学習データの性別ごとのデフォルト，非デフォルトの分布を確認して特定の性別のデフォルトが高いデータがそろっていないかを確認する．（付録 E（１）性別を参照） ・モデルの公平性指標(Equalized Odds Difference）の計算値が 0 に近いスコアになっているかを確認する．（付録 D システムの基本設計情報を参照） ・上記 2 点の確認内容が満足されていない場合は，判定結果に対して，Fairlearn（Microsoft）等の補正アルゴリズムでモデルが出力する判定結果の公平性のバイアスを軽減する処理をしているかを確認する．（付録 D システムの基本設計情報を参照）
	補足説明	Fairlearn が提供するアルゴリズムにより，機械学習の推論結果の不公平性を軽減できる．公平性に関する補正処理がない一般のアルゴリズム（例: LightGBM）に比べ，補正ありのアルゴリズム（ThresholdOptimizer/ GridSearch）では，正答率は低下するが公平性は改善される． 
8	要件	機動的なモデルの再学習
	確認事項	・再学習の時間は，運用で許容できる時間以内であるか．
	確認手段	・機械学習の学習時間，学習方式，モデル構築等の条件をみて，運用フェーズでの再学習が著しく長くなるような懸念がないかを確認する．（付録 D システムの基本設計情報および付録 E（６）モデル性能を参照）
9	要件	管轄省庁のガイドラインに準拠
	確認事項	・収入が低い世代の人に対してのクレジット額が高くなっていないか．
	確認手段	・年代ごとの与信のデータ分布を確認して，収入が低い世代に多額の与信額が割り当てられていないかを確認する．※経済産業省「改正割賦販売法」の過剰与信防止義務の確認（付録 D を参照）
10	要件	事故発生時の解析の容易性
	確認事項	・学習データ，検証データおよびモデルの学習履歴の検証が必要な局面で，いつでも参照できるように記録が残されているか．
	確認手段	・機械学習の学習時および検証時に使用した学習データセット，検証データセットが保管されており，必要なときに参照できるかを確認する．（付録 D システムの基本設計情報を参照）

付録 G　仮説検証を目的とした実験に関する被験者の回答データ

※論文 5.2 節「RQ に関する評価結果」のデータ（t 検定の対象サンプル）

表 G-1 被験者回答精度（Ｉ群$\sum_i A_i \cdot Cup(g_{i\omega})$）

No.	所属企業	機械学習技術の実務経験*1(年)	Ｉ群被験者	条件 a	条件 b	条件 c
1	製造業	0	A 社 QA 担当	11.5	17.5	41.7
2	製造業	0	B 社 QA 担当 1	9.4	16.1	35.9
3	製造業	0	B 社 QA 担当 2	7.1	14.4	37.5
4	製造業	0	B 社 QA 担当 3	11.5	19.2	41.2
5	製造業	0	C 社 マネージャ	23.4	22.5	39.2
6	製造業	0	C 社 QA 担当	9.2	18.6	32.6
7	製造業	0	D 社 開発担当 1	5.0	17.5	33.6
8	製造業	0	D 社 開発担当 2	11.2	12.9	24.3
9	製造業	0	D 社 開発担当 3	8.9	11.2	22.3
10	製造業	1	D 社 開発担当 4	6.8	11.1	20.2
11	製造業	0	D 社 開発担当 5	8.9	0.0	—
12	製造業	0	D 社 研究員 1	6.5	7.4	19.3
13	製造業	0	D 社 開発担当 6	12.8	24.7	33.1
14	製造業	0	D 社 開発担当 7	7.7	3.2	—
15	製造業	0	D 社 開発担当 8	11.2	6.8	31.0
16	情報・通信	0	E 社 QA 担当	15.5	24.7	29.7
17	情報・通信	0	F 社 開発担当	2.9	7.6	14.7
18	情報・通信	0	G 社 開発担当 1	11.7	11.7	33.7
19	情報・通信	0	G 社 開発担当 2	4.7	10.9	12.3
20	情報・通信	1	G 社 開発担当 3	7.1	13.8	22.6
21	情報・通信	0	G 社 その他 1	7.3	11.5	15.5
22	情報・通信	0.5	G 社 その他 2	4.7	12.6	14.1
23	情報・通信	0	G 社 リーダ 1	5.3	6.7	7.6
24	情報・通信	0.5	G 社 開発担当 4	6.8	11.5	16.4
25	情報・通信	0	G 社 開発担当 5	7.7	14.9	20.7
			平均	9.0	13.2	26.1
			分散	17.8	37.9	102.9
			要件充足率	18.9%	27.6%	54.7%
			サンプル数	25	25	23

【凡例】—：実験データ回収不可

(*1) 機械学習の実務経験がない被験者，若しくは 1 年以下で非研究員の被験者はＩ群に該当

表 G-2 被験者回答精度（Ⅱ群$\sum_i A_i \cdot Cup(g_{i\omega})$）

No.	所属企業	機械学習技術の実務経験*2(年)	Ⅱ群被験者	条件 a	条件 b	条件 c
1	製造業	5	A 社 マネージャ	19.9	27.1	38.5
2	製造業	2	H 社 リーダ	32.1	27.8	42.0
3	情報・通信	3	I 社 リーダ	16.0	16.8	40.3
4	製造業	12	J 社 マネージャ	10.4	23.1	34.0
5	製造業	15	K 社 マネージャ	18.6	17.6	29.0
6	製造業	2	D 社 開発担当 9	19.8	17.9	31.8
7	製造業	3	D 社 開発担当 10	11.3	13.5	30.9
8	製造業	0.5	D 社 研究員 2	19.3	21.2	33.4
9	金融	2	L 社 研究員	16.8	—	—
10	情報・通信	3	G 社 開発担当 6	14.5	20.4	24.8
11	情報・通信	3	G 社 リーダ 2	14.4	22.6	10.9
12	情報・通信	2	G 社 リーダ 3	11.0	11.5	21.6
13	製造業	10	M 社 リーダ	14.0	16.4	29.6
			平均	16.8	19.7	30.6
			分散	32.3	26.4	73.9
			要件充足率	35.2%	41.3%	64.2%
			サンプル数	13	12	12

【凡例】—：実験データ回収不可

(*2) 機械学習の実務経験が 2 年以上の被験者，若しくは 0.5 年以上で研究員の被験者はⅡ群に該当

以上