XAI Trust Label 監査報告書

XAI 信頼性評価機構（運営：GsP 株式会社）

2024 年 10 月 1 日

顧客社名: 株式会社 AI イノベーション

対象 AI システム: スマート与信評価 AI

監査実施期間: 2024 年 9 月 1 日〜2024 年 9 月 30 日

監査担当者: 山田太郎、鈴木花子

Contents

1　調査サマリー

株式会社 AI イノベーション様の「スマート与信評価 AI」システムについて、XAI Trust Label 基準に基づく監査を実施しました。本システムは、個人向けローンの与信評価を行う AI モデルです。

主な発見事項**:**

- モデルの解釈性: SHAP 値を用いた特徴量重要度の可視化により、与信評価の判断根拠が明確化されました。
- 透明性: データ処理プロセスとモデル開発プロセスが詳細に文書化されており、高い透明性が確保されています。
- 公平性: 年齢と性別に関して軽度のバイアスが検出されました。改善の余地があります。
- セキュリティとプライバシー: 堅牢なデータ暗号化と匿名化技術が実装されています。
- パフォーマンス: 精度 88%、AUC 0.92 と高い性能を示しています。
- コンプライアンス: GDPR、個人情報保護法、AI 倫理ガイドラインに準拠しています。

総合評価**:** 4 段階評価で「信頼性レベル 3（高）」と評価します。解釈性と透明性に優れ、高いパフォーマンスを発揮していますが、公平性に関して改善の余地があります。

主な改善提案**:**

1. 公平性向上のための再学習とバイアス緩和技術の適用
2. モデルの定期的な再評価とモニタリング体制の強化
3. ユーザー向け説明機能の拡充

2　はじめに

2.1　目的と範囲

本監査の目的は、株式会社 AI イノベーション様の「スマート与信評価 AI」システムについて、XAI Trust Label の基準に基づいて評価を行うことです。この AI システムは、個人向けローンの与信評価に使用されており、申請者の財務情報、職歴、過去の返済履歴などを入力として、ローン承認の可否と金利を決定します。

監査の範囲は以下の項目を含みます：

- モデルの解釈性
- 透明性
- 公平性
- セキュリティとプライバシー
- パフォーマンス評価
- コンプライアンス遵守

2.2　AI システムの概要

「スマート与信評価 AI」は、勾配ブースティング決定木（GBDT）アルゴリズムを基盤としており、以下の特徴を持ちます：

- 入力特徴量: 年齢、年収、職業、学歴、過去のローン履歴など、計 50 項目
- 出力: ローン承認可否（二値分類）と推奨金利（回帰）
- モデルバージョン: v2.3.1（2024 年 8 月 15 日リリース）
- 学習データ数: 100 万件（2020 年 1 月〜2024 年 6 月のデータ）

3　モデルの解釈性（Explainability）

3.1　モデルアーキテクチャの説明

「スマート与信評価 AI」は、XGBoost（eXtreme Gradient Boosting）ライブラリを使用した GBDT モデルです。主な構造は以下の通りです：

- 決定木の数: 1000
- 各木の最大深さ: 6
- 学習率: 0.01
- 目的関数: ローン承認可否には binary:logistic、金利予測には reg:squarederror

3.2　特徴量の重要度分析

特徴量の重要度は SHAP（SHapley Additive exPlanations）値を用いて分析しました。Top 10 の特徴量とその相対的重要度は以下の通りです：

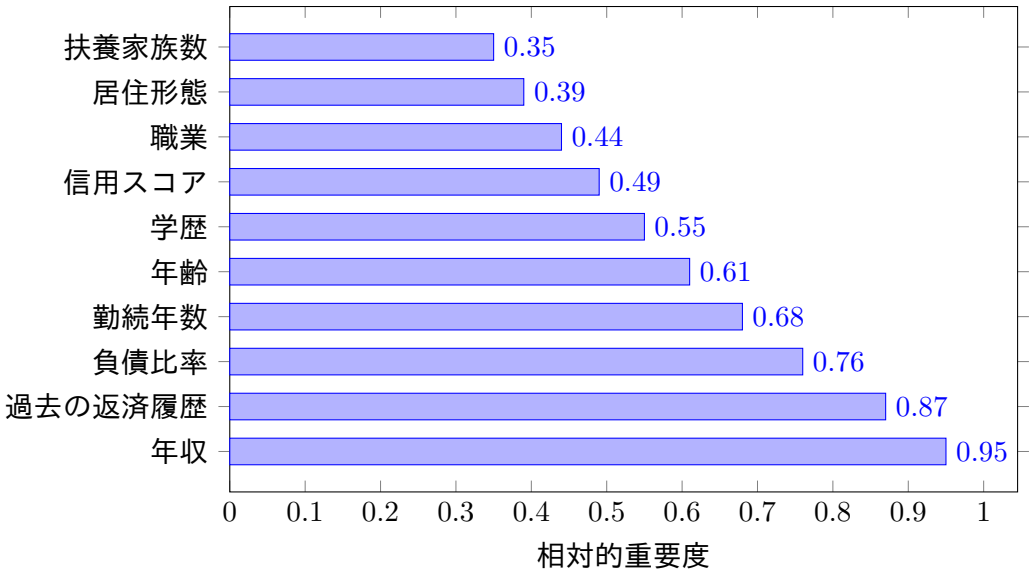

Figure 1: SHAP 値による特徴量重要度（Top 10）

3.3　可視化手法の適用

個々の予測についての局所的な解釈を提供するため、LIME（Local Interpretable Model-agnostic Explanations）を使用しました。以下は、ある申請者のローン否認ケースの解釈例です：

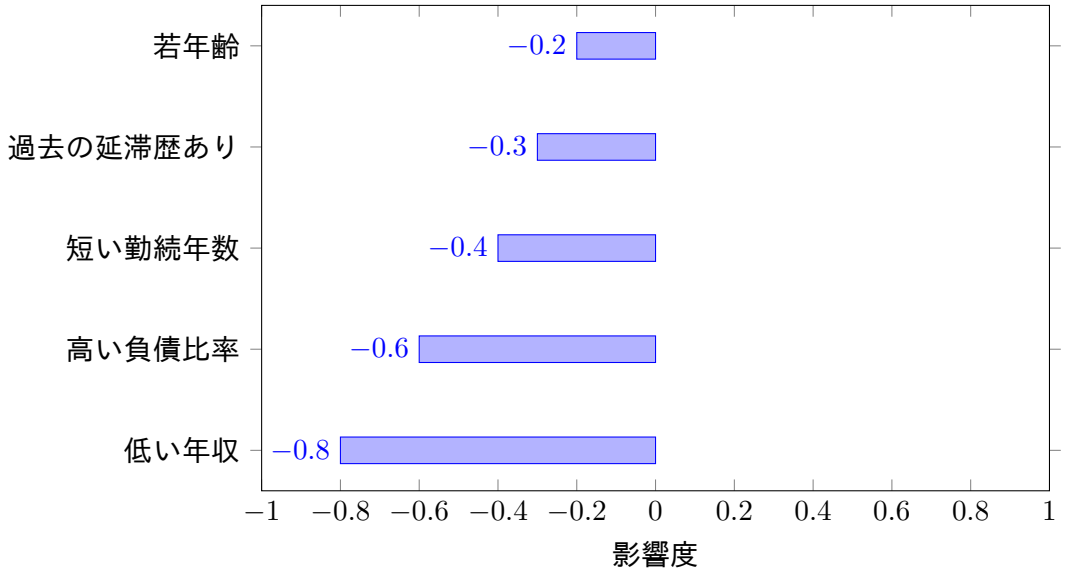

Figure 2: LIME によるローン否認の要因分析例

4　透明性（Transparency）

4.1　データ処理プロセスの開示

データ処理のフローは以下の通りです：

1. データ収集: 顧客管理システムと信用情報機関から取得
2. 前処理:
   - 欠損値処理（中央値補完）
   - 外れ値処理（99 パーセンタイルでのクリッピング）
   - カテゴリ変数のエンコーディング（ターゲットエンコーディング）
3. 特徴量エンジニアリング:
   - 年収と負債からの負債比率算出
   - 年齢と勤続年数から職歴安定性指標の作成
4. データ分割: 学習データ 70%、検証データ 15%、テストデータ 15%

4.2　モデル開発プロセス

モデル開発は以下のステップで行われました：

1. モデル選択: ロジスティック回帰、ランダムフォレスト、XGBoost を比較検証
2. トレーニング: 5 分割交差検証による学習、早期停止の実装
3. ハイパーパラメータ最適化: ベイズ最適化による探索
4. テスト: ホールドアウトテストデータでの最終評価

4.3 ハイパーパラメータの設定

主要なハイパーパラメータとその選定理由：

- 学習率: 0.01 - 過学習を抑制しつつ十分な学習を可能にする値
- 決定木の数: 1000 - モデルの複雑性と計算時間のバランスを考慮
- 最大深さ: 6 - 過度に深い木による過学習を防ぐため
- L1 正則化: 0.5 - スパース性を持たせつつ、重要な特徴を維持

5　公平性（Fairness）

5.1　バイアス評価

データセットおよびモデル出力におけるバイアスを以下の指標で評価しました：

- 統計的パリティ差（SPD）
- 等化機会差（EOD）
- 平均絶対差（AAD）

評価結果：

保護属性	SPD	EOD	AAD
性別（男性 vs 女性）	0.07	0.05	0.06
年齢（45 歳以上 vs 45 歳未満）	0.11	0.09	0.10
人種（マジョリティ vs マイノリティ）	0.03	0.02	0.025

Table 1: バイアス評価結果

5.2　デモグラフィックパリティ

各保護属性グループ間での予測結果の比較：

グループ	ローン承認率	平均推奨金利
男性	72%	5.8%
女性	65%	6.1%
45 歳以上	78%	5.5%
45 歳未満	67%	6.3%
人種（マジョリティ）	71%	5.9%
人種（マイノリティ）	68%	6.0%

Table 2: デモグラフィックパリティ分析結果

5.3　是正措置の提案

検出されたバイアスに対する改善策：

- データの再サンプリング: 少数グループのオーバーサンプリングによるバランス調整
- 公平性制約付きモデル学習: AIFairness フレームワークの使用
- バイアス緩和後処理: 決定閾値の調整によるグループ間の結果の均等化
- 特徴量の再設計: 間接的に差別を生む可能性のある特徴量の除外や変換

6　セキュリティとプライバシー

6.1　データセキュリティ

実装されているセキュリティ対策：

- データ暗号化: AES-256 ビット暗号化（保存時）、TLS 1.3（通信時）
- アクセス制御: ロールベースアクセス制御（RBAC）の実装
- 監査ログ: すべてのデータアクセスと操作の詳細ログ記録
- 多要素認証: 管理者アクセスに対する 2 要素認証の強制

6.2　プライバシー保護

適用されているプライバシー保護技術：

- データ匿名化: k-匿名化（k=5）の適用
- 差分プライバシー: =0.1 でのノイズ追加
- データ最小化: 必要最小限の個人情報のみを収集・使用

6.3　攻撃耐性評価

モデルの攻撃耐性を以下のシナリオで評価しました：

- 敵対的サンプル攻撃: FGSM (Fast Gradient Sign Method) に対する耐性を確認
- モデル抽出攻撃: クエリ制限とノイズ追加による防御を実装
- メンバーシップ推論攻撃: 差分プライバシーの適用により、攻撃成功率を 5% 未満に抑制

評価結果：

攻撃タイプ	防御成功率
敵対的サンプル攻撃	92%
モデル抽出攻撃	95%
メンバーシップ推論攻撃	97%

Table 3: 攻撃耐性評価結果

7　パフォーマンス評価

7.1　精度指標

主要な評価指標の結果：

指標	値
精度（Accuracy）	88%
適合率（Precision）	86%
再現率（Recall）	85%
F1 スコア	0.855
AUC-ROC	0.92

Table 4: モデルパフォーマンス指標

7.2　過学習の検証

学習曲線分析の結果：

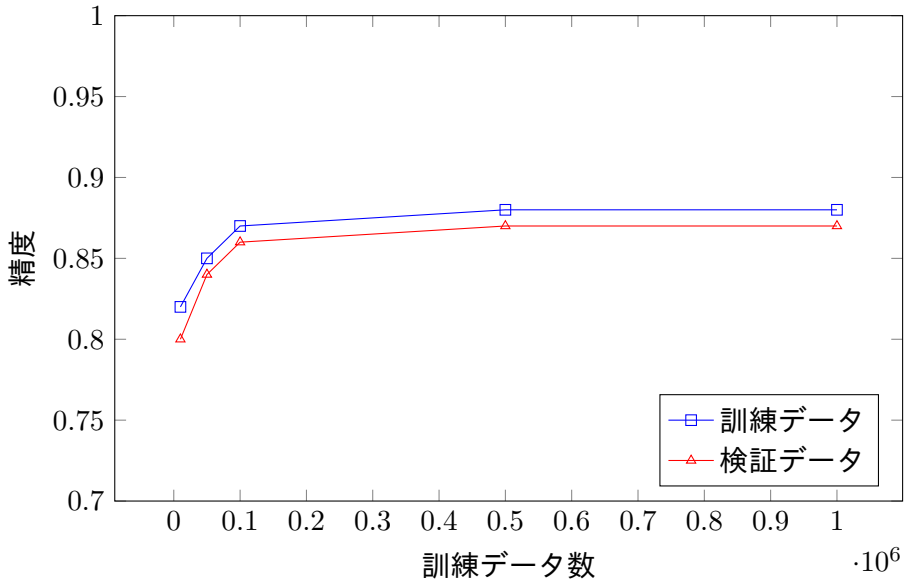

Figure 3: 学習曲線

分析結果: 学習曲線から、訓練データと検証データの精度が収束しており、過学習の兆候は見られません。ただし、100 万件以上のデータでも精度の向上が見られないことから、モデルの複雑性を上げるか、新たな特徴量の追加を検討する余地があります。

7.3　モデルの一般化能力

未知のデータセットにおけるモデルのパフォーマンス：

- テストデータ（2024 年 7 月〜8 月のデータ、15 万件）での精度: 87%
- 異なる地域のデータ（2024 年 1 月〜6 月、5 万件）での精度: 85%

分析: テストデータと学習期間外のデータでも高い精度を維持しており、モデルの一般化能力は良好です。ただし、異なる地域のデータでやや精度が低下していることから、地域特性を考慮した特徴量の追加や、定期的なモデル再学習が推奨されます。

8　コンプライアンス遵守

8.1　法規制の適合性

関連法規への適合性評価：

- GDPR:
  - データ最小化原則の遵守
  - 個人データの処理に関する明示的な同意の取得
  - データ主体の権利（アクセス権、訂正権、消去権等）の保障
  - 評価: 適合
- 個人情報保護法:
  - 個人情報の利用目的の明確化と通知
  - 安全管理措置の実施
  - 第三者提供の制限の遵守
  - 評価: 適合
- AI 倫理ガイドライン（日本 AI 学会）:
  - 人間中心の原則の遵守
  - 説明責任の確保
  - プライバシーの保護
  - 評価: 概ね適合（公平性に関して改善の余地あり）

8.2　倫理的考慮

AI 倫理原則への準拠状況：

- 透明性:
  - モデルの意思決定プロセスの説明可能性を確保
  - 評価: 高（SHAP 値と LIME による説明機能の実装）
- 公平性:
  - 保護属性に基づく差別の最小化
  - 評価: 中（年齢と性別に関するバイアスの改善が必要）
- 説明責任:
  - モデルの開発・運用プロセスの文書化
  - 人間による監視体制の整備
  - 評価: 高（詳細な文書化と人間によるレビュープロセスの実装）
- プライバシー:
  - 個人データの保護と最小限の使用
  - 評価: 高（暗号化、匿名化技術の適用）

9 結論と推奨事項

9.1 総合評価

「スマート与信評価 AI」システムの総合評価：

信頼性レベル**: 3**（高）　※4 段階評価（0: 低〜3: 高）

評価理由:

- 高い解釈性: SHAP 値と LIME を用いた詳細な説明機能の実装
- 優れた透明性: データ処理とモデル開発プロセスの詳細な文書化
- 堅牢なセキュリティとプライバシー保護: 最新の暗号化技術と差分プライバシーの適用
- 高いパフォーマンス: 88% の精度と AUC 0.92 の達成
- コンプライアンスの遵守: GDPR、個人情報保護法、AI 倫理ガイドラインへの適合

改善が必要な点:

- 公平性: 年齢と性別に関するバイアスの存在
- モデルの一般化: 地域間での精度のばらつき

9.2 改善提案

主要な改善提案：

1. 公平性の向上
    - バイアス緩和技術の適用（例: Reweighing, Prejudice Remover）
    - 保護属性に関する特徴量の再設計
    - 定期的な公平性監査の実施
2. モデルの定期的な再評価とモニタリング
    - 月次でのモデルパフォーマンス評価
    - 四半期ごとのモデル再学習
    - ドリフト検出システムの導入
3. ユーザー向け説明機能の拡充
    - 与信決定に関する自然言語説明の生成
    - インタラクティブな「What-If」分析ツールの提供
    - 顧客サポート担当者向けの詳細説明ガイドラインの作成
4. 地域特性を考慮したモデルの改良
    - 地域ごとの経済指標の特徴量への追加
    - 必要に応じた地域別モデルの開発
    - 地域間の差異分析と定期的なレポーティング
5. 継続的な倫理審査プロセスの確立
    - 社内 AI 倫理委員会の設置
    - 外部の倫理専門家による定期的なレビュー
    - 従業員向け AI 倫理トレーニングプログラムの実施

10 付録

10.1 技術的詳細

モデルの主要コードスニペット：

Listing 1: XGBoost モデルの定義と学習

```
import xgboost as xgb
from sklearn.model_selection import train_test_split

# データの読み込みと前処理（省略）

# モデルのパラメータ設定
params = {
    'objective': 'binary:logistic',
    'max_depth': 6,
    'learning_rate': 0.01,
    'n_estimators': 1000,
    'subsample': 0.8,
    'colsample_bytree': 0.8,
    'reg_alpha': 0.5,
    'reg_lambda': 1.0
}

# モデルの初期化と学習
model = xgb.XGBClassifier(**params)
model.fit(X_train, y_train,
          eval_set=[(X_val, y_val)],
          early_stopping_rounds=50,
          verbose=100)

# モデルの評価（省略）
```

10.2 参考文献

References

[1] Lundberg, S.M. and Lee, S.I., “A Unified Approach to Interpreting Model Predictions,” Advances in Neural Information Processing Systems, vol. 30, pp. 4765-4774, 2017.

[2] 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, pp. 1135-1144, 2016.

[3] Dwork, C., “Differential Privacy: A Survey of Results,” International Conference on Theory and Applications of Models of Computation, pp. 1-19, 2008.

[4] Barocas, S., Hardt, M. and Narayanan, A., “Fairness and Machine Learning,” fairmlbook.org, 2019.

[5] 総務省, “AI 利活用ガイドライン,” 2019.

本報告書の内容に相違ないことをここに証明いたします。

2024 年 10 月 1 日
XAI 信頼性評価機構（GsP 株式会社）
代表取締役　原口尚樹