2025-12-25 MTG(阿曽村): フィッシング検知の二層パイプラインと確信度設計

開会と導入の雰囲気

「いけてる気がする。」[Speaker 1]の軽い一言から会は柔らかく始まり、[Speaker 2]が「大丈夫です」と応じて、落ち着いた空気のまま技術的な報告へと移った。全体として、互いに敬意を払いながらも率直に課題へ向き合う、前向きで実務的なムードが終始続いた。

研究の前提確認と目的の再定義

[Speaker 1]がまず、12月4日の打ち合わせ内容を再確認した。「XGBoostのスレッショルドを微調整しながら、第2層のエージェントに回すSNS（≒URL/サイト）数を調整してみよう」という宿題を受けて進めてきたという。研究の骨子は、フィッシングサイトの証明書情報を用い、二層構成で検知すること。第1層はXGBoostで正常/フィッシングを粗く選別し、第2層でAIエージェントが精査する「二段構え」のパイプラインだ。

ここで[Speaker 2]が確認を挟む。「最初のパイプラインで判定が済むのか、それとも全部やってエージェントも含めた最終結果なのか。」これに対し[Speaker 1]は「本日の報告は第1層（XGBoost）に関するものが中心で、続きの報告もある」と明示。議論の焦点が第1層のしきい値調整にあることが、参加者間で共有された。

データ条件と初期結果の提示

[Speaker 1]は前提データを整理した。学習に用いたデータセットからテスト用に切り出した規模は「12万8000通り」。しきい値0.45の場合の結果として、誤分類に関係する数値を提示したが、説明は断片的で「二度足が3843」「1664」という言葉が続き、混乱が生じた。元々は「閾値0.5で見逃し（False Negative）が4188、欠け値（恐らくFalse Positive）が1312」でバランス重視の設定だったが、しきい値をずらして探索し、0.45付近がよさそうだと感じた一方、見逃しを大胆に減らすため0.2まで下げると別の副作用（誤検知増加）が大きいことを、過去の分布図とともに言及した。

「0.2〜0.4あたりがバランスは良い」という一次評価に対して、[Speaker 2]は「全体の設計意図は理解した。第1層はXGBoost、第2層でエージェント。いま話している結果は第1層だけの話で合っているか」と確認し、[Speaker 1]が「入口として第1層の話」と確定した。

論点転換：第2層に渡すデータ定義と「確信度」の問題

ここで[Speaker 2]が本質的な問いを投げる。「第2層が処理すべきものは具体的にどういう問題か。『フォールスネガティブを減らす』は理想だが、現実の運用では正解は未知であり、『確信を持てない』とは何を指すのか、定義と数が重要だ。」[Speaker 1]は「確信度の定義はこれから考える」と認め、核心課題が明確化した。

[Speaker 2]は続ける。「重要なのは第1層から第2層に渡す『確信のない（曖昧）サンプル』をどう定義し、どれだけ減らせるかだ。」これにより議論は、単なる閾値調整から「確信度駆動の2層連携設計」へとシフトした。

第2層（エージェント）の単体性能と現実分布

[Speaker 1]は、第2層エージェントの単体テストについて触れた。「False Negativeだけを渡した場合は95%当選（＝正しく検出）できるようになった。」また、正常データを混ぜたバランスデータ（半々）でも「まあまあ選り分けられた」と述べ、クロストークの指標では「保険値は24点」と表現した（数値の意味は要確認：[fill in the blank]）。[Speaker 2]は「エージェント単体の精度は十分。ただ鍵は第1層と第2層をどう繋げるか」と指摘し、設計の重要性を再度強調した。

さらに[Speaker 2]は現実のクラス分布（正規が多数、フィッシングが少数）に言及。「バランスデータではなく、実分布（例：正規8割、フィッシング2割）で評価した場合、精度は下がる可能性があるため、そこも試すべき」と助言。これに対し[Speaker 1]も同意し、過去の不正送金検知の経験に引き寄せて「少数派検知の評価設計」を再認識した。

「自信を持って間違える」ケースへの対処

中盤の重要な論点は、「モデルが自信を持って誤判定する」ケース（高確信度のFalse Positive/False Negative）。[Speaker 2]は「0や1に張り付く高確信度サンプルは、本来『絶対正しい』であってほしい。中間（グレー）はグレーのままにすべき。今の訓練は曖昧なものを無理に0/1へ押し込んでいる可能性がある。」と問題設定を再構成。

[Speaker 1]も「自信を持って間違えている箇所は数値で抽出可能で、特徴を見てチューニングができる」と応じ、具体的な改善の方向性が合意された。

設計提案：グレー判定を許容する第1層

議論は、「第1層がグレー（不確実）判定を許容する」方向で収束していく。[Speaker 2]は、二値分類の訓練方針を修正する発想を示した。「0/1に無理に寄せず、中間を許容し、確信度が低いサンプルは『保留（グレー）』とし、第2層で精査する。確信度が高い0/1はほぼ正しいことを保証する。」この戦略では、第1層の二値性能（Precision/Recall）は多少落ちてもよい。代わりに誤った高確信度判定を減らし、中間（グレー）を第2層に渡す設計が合理的だ。

[Speaker 1]は「XGBoostにこだわらず、グレーを扱える設計にしたい」という前向きな姿勢を示し、[Speaker 2]も「三値分類（0=悪性、1=正常、2=グレー）なども選択肢。

モデルもXGBoostに限らない」と柔軟性を認めた。「最近はLLMの支援も有効。Claude、Gemini、CLIベースの支援も設計段階で役に立つ」との実務的助言が重ねられた。

技術的合意と今後の検証方針

やり取りの末、両者で以下の技術的ポイントが合意に達した。

- 「確信度の定義」を明確化し、第1層でグレー判定を許容することで、第2層に渡す対象を「確信不足のもの」に限定する。
- 「自信を持って間違える」高確信度誤判定を特定し、特徴分析により第1層をチューニングして削減する。
- 評価は、実分布（正規多数）でも検証し、Precision/Recallなどに加え、グレー判定率、グレー→第2層での最終正解率も含める。
- モデルはXGBoostに固執せず、三値や確信度出力を前提にした別手法も検討する。

この過程で[Speaker 1]は「できる気がしてきた」と何度も口にし、議論が具体的な突破口へ近づいたことが会話のトーンからも感じられた。

人的側面：時間制約とモチベーション

後半、研究の進捗と個人的事情が共有された。[Speaker 1]は「年単位で全振りして進めてきたが、進まない焦りがある。3月の卒業で時間切れの懸念。」と打ち明ける。[Speaker 2]は「社会人の時間制約は当然。博士取得権利は卒業後数年あるはず。焦り過ぎず、今は設計を明確化して進めるべき。」と励ましつつ現実的な見取り図を示した。エージェントは「成長してきて自信を持てる」との[Speaker 1]の言葉に対しても、[Speaker 2]は肯定し、「第1層の再設計でグレーの精度を上げ、グレー以外はほぼ正しい状態へ」—という、論文化に耐える問題設定の明確さを強調した。

ツール利用についても率直な対話があった。[Speaker 1]は「Claudeを解約したが再契約を検討」と述べ、[Speaker 2]は「LLMを設計支援に使うのは有用。コードの自動生成も、理解していれば問題ない」と後押しした。

締めくくりの挨拶と年末年始の計画

終盤は互いの労をねぎらい、年末年始の挨拶が交わされた。「良いお年を」「メリークリスマス」「また来年もよろしく」。[Speaker 1]は「結果をチャットで逐次共有する」と約束し、[Speaker 2]は「年末年始は集中できる時期、適宜進めてください」と応じた。全体として温かいがプロフェッショナルな空気で閉じた。

決定事項

- 第1層の設計方針を「グレー判定を許容する」方向へ切り替える（XGBoostに固執しない）。

- 高確信度の誤判定（自信を持って間違える）を特定・分析し、特徴ベースで第1層のチューニングを実施。
- 評価設計を、バランスデータに加え、実分布（例：正規多数、フィッシング少数）での検証へ拡張。
- 第2層（エージェント）への入力は「確信度不足のグレー判定」に限定する運用を目指す。

未解決事項・要定義

- 「確信度」の具体的定義と算出方法（例：確率閾値、予測分布のエントロピー、キャリブレーション指標など）[fill in the blank]
- 三値分類（0/1/グレー）におけるラベル設計・学習戦略（損失関数、サンプリング、コスト重み付け）[fill in the blank]
- 「保険値24点」の評価指標の意味と算出根拠[fill in the blank]
- 実分布の想定（例：正規:フィッシング=8:2の妥当性、運用ドメインでの真の比率）[fill in the blank]

アクションアイテム

- [Speaker 1]：第1層における「確信度不足（グレー）」を定義し、グレー判定を許容するモデル案を複数試作（XGBoostの閾値運用に加え、三値分類や別モデルを比較）。期限：[fill in the blank]
- [Speaker 1]：高確信度誤判定の抽出と特徴分析（誤判定クラスタの可視化、特徴重要度、SHAP/Permutationの検討）。期限：[fill in the blank]
- [Speaker 1]：評価プロトコルを拡張（実分布でのPrecision/Recall、グレー率、グレー→第2層の最終精度、処理コストの測定）。期限：[fill in the blank]
- [Speaker 2]：モデル選定・訓練戦略（グレー許容のロス設計やキャリブレーション手法）に関する参考資料・アルゴリズムの提案。期限：[fill in the blank]
- [Speaker 1]：LLM支援環境の再整備（Claude/Geminiの契約、CLIワークフロー構築）とプロトタイピング着手。期限：[fill in the blank]
- 両者：年末年始期間中に進捗をチャットで随時共有、次回打ち合わせ日程の調整。期限：[fill in the blank]

次のステップとフォローアップ

次の重点は、第1層を「二値の押し込みから、確信に基づく三相（確信0/確信1/グレー）設計」へと転換し、グレーのみ第2層に渡すパイプラインの成立を確認すること。並行して、「自信を持って間違える」事例を減らすための特徴分析とチューニングを実施する。評価は、実分布を踏まえた指標で再設計し、コスト（第2層の計算負荷）と精度のトレードオフを定量化する。設計・実装にはLLMを活用し、モデル選定や損失設計に関する代替案を幅広く試す。

総括

本会議は、単なるしきい値探索から一歩進み、「確信度駆動の二層連携」という明確な設計思想を共有できた。第1層の役割を「絶対に正しい0/1のみを確信を持って返し、曖昧はグレーへ」と再定義し、第2層エージェントがグレーに集中する流れが合意されたことで、全体の整合性と説明可能性が高まった。人的・時間的制約の共有により、現実的な進め方と励ましが交わされ、年末年始に向けた実務的な前進が期待される。「できる気がしてきた」という[Speaker 1]の言葉どおり、次回の報告で具体的な成果（グレーの定義、誤判定の削減、評価指標の整備）が示されることが、ステークホルダーにとっての最大の関心事となるだろう。