修士論文

開発途上国のデータ駆動型信用スコアリングモデルの解釈可能化

―ベトナム海運商業銀行の顧客データの分析―

松永　智行

主指導教員　　Dam Hieu Chi 教授

北陸先端科学技術大学院大学
先端科学技術研究科
（知識科学）

令和 3 年 3 月

Abstract

Financial services are important in developing countries, and the lack of them is one of the major obstacles for many poor people to improve their living standards. Recently, a number of methods and businesses including FinTech have emerged to address this problem by combining machine learning technology with data such as digital footprints, which had not been used to determine their creditworthiness. However, there is a problem with these methods. That is the anxiety over whether we can really trust the information on creditworthiness (credit score) generated by such new methods. This concern is not limited to the financial sector, and now that AI technology has begun to spread, it is being discussed in various fields as an issue of interpretability in AI (machine learning).

In this study, we examined the need for, and problems of, interpretability in machine learning, which has not been studied in detail in Japan. The insights therefrom were applied to credit scoring model research in the field of finance. In addition, we applied principal component analysis (PCA), clustering, and Association Rule Mining to a data set provided by the Vietnam Maritime Commercial Joint Stock Bank that linked customers' credit scores to their SNS posts. These analyses confirmed some tendencies that existed between the content of SNS posts and credit scores, and contributed, albeit partially, to improving the interpretability of the credit scoring model.

A major question remains: "To what extent should we be able to interpret (explain) the data so that people will trust or accept it? " Our idea of a good explanation is extremely vague and complex due to the problems of truthfulness and confirmation bias, and the point at which an explanation is judged to be good may be affected by people's emotions and interests. Therefore, if the compromise point of "how much explanation is enough" is not appropriate, the cost of credit screening will increase and the premise of the ability to solve financial problems will be weakened. It should be noted that the expansion of FinTech businesses in developing countries was fostered by the looser legal regulations compared to developed countries.

目次

第 1 章 序論 1

1.1 研究背景 1

1.1.1 開発途上国における金融課題 1

1.1.2 フィンテックによる金融課題の解決 2

1.1.3 新たな課題とモデルの解釈可能性 3

1.2 研究目的 4

1.3 知識科学的意義 5

1.4 本論文の構成 5

第 2 章 関連研究 7

2.1 解釈可能な AI(XAI)について 7

2.1.1 社会情勢・背景 7

2.1.2 研究界の動向 9

2.1.3 解釈可能性の必要性 10

2.1.4 解釈可能性を必要としない場合 12

2.1.5 解釈手法の分類 14

2.1.6 人間にとって良い説明とは 16

2.1.7 現状の課題 20

第 3 章 分析手法とデータ 22

3.1 信用スコアリングモデルにおける解釈可能性 22

3.2 使用するデータ 23

3.3 分析手法について 25

3.3.1 クラスタリング 25

3.3.2 相関ルール分析 27

第 4 章 結果と考察 29

4.1 クラスタリングの結果 29

4.2 相関ルール分析の結果 33

4.3 まとめと考察 41

第 5 章 総括と今後の課題 43

5.1 総括 43

5.2 今後の課題 45

参考文献 46

謝辞 48

表目次

表 3.1 SNS への投稿内容のデータの例..... 25

表 4.1 各主成分の固有ベクトル..... 30

表 4.2 分割数を 6 としたクラスタリングの結果..... 30

表 4.3 分割数を 8 としたクラスタリングの結果 1..... 32

表 4.4 分割数を 8 としたクラスタリングの結果 2..... 33

表 4.5 各クラスターへ設定した最低支持度と得られたルールの数..... 35

表 4.6 クラスター4 の結果 36

表 4.7 クラスター5 の結果 37

表 4.8 クラスター6 の結果 37

表 4.9 信頼度 A のみの結果..... 40

表 4.10 信頼度 C のみの結果..... 40

第1章 序論

1.1 研究背景

1.1.1 開発途上国における金融課題[1]

貧困削減は世界の課題の一つであり、また開発経済学では学問領域全体にまたがるテーマであり続けている。それらの中でも、近年重要視されているのが開発途上国における金融サービスの充足度であり、貧困層が生活改善を行う上で金融サービスへのアクセスの欠如が、大きな障害の一つとなっている。そして、その金融サービスの欠如の原因の一つが、開発途上国では融資する側の金融機関からすると、低所得者を中心として本人確認や信用度合いの判断が困難な層が少なからず存在することにある。金融機関は返済能力のある対象者の選抜を行う必要があるが、それ自体が高コストとなるため結果的に金融サービスの提供も困難になるのである。

銀行をはじめとするフォーマルな金融サービスから排除された人々は、金利の高いインフォーマルな金融機関か、限られた額しか貸し出せない友人や家族に頼ることになる。また、中小零細企業や個人事業主が、必要なときに融資を受けられないことは、事業の成長機会を逃したり、緊急時に乗り切れないなど、個

[1] 1.1.1 および 1.1.2 は主に [1]および [2]を参考とした。

人以上に悪影響が大きい。

1.1.2 フィンテックによる金融課題の解決

以上の事例を含む開発途上国における金融課題の解決策の一つとして、近年ではフィンテック、すなわち金融と IT とを融合させた技術とビジネスが期待されている。これは融資と信用情報に関わる課題の解決も例外でない。東南アジア地域におけるインターネットとスマートフォンの普及は、利用者が金融サービスを受けるためにわざわざ銀行の支店や ATM に赴く手間を省き、モバイル決済を可能にした。さらに、そのモバイル決済に使用する資金としては電子マネーやビットコインが利用できるため銀行口座やクレジットカードの保有も必要ない。一方で金融サービスの提供側は、インターネットの特性である物理的な距離に縛られないことを利用して本人確認などを行うことによって以前より格段に低いコストで顧客情報を取得できるようになった。また、金融サービスの提供者がアクセスできる情報はデジタル・フットプリント（デジタル上の足跡）にまで拡大しており、これまで信用度合いの測定では使われてこなかった情報を大量に取得し、それらと機械学習技術を組み合わせることで低コストかつ迅速に信用度合いの測定を行う手法が出現している。

この手法の具体例としてシンガポールに本社を置く Lenddo がある。同社は

銀行、マイクロファイナンス機関、クレジットカード発行会社などに個人の信用スコア「LenddoScore」および本人確認サービスの「Lenddo Verification」を提供している。Lenddo 社が収集・解析するデータは、スマートフォンから得られる個人データ、SNS 上から得られる個人データ、心理測定で得られる個人データなどである。本研究に関わりが深い SNS 上のデータで具体例を挙げると、FaceBook、Linkedin、Twitter 等のアクセス先、アクセス頻度、友達の数、メッセージの投稿内容、等である。以上のものと信用機関にデータがある場合はそれらも組み入れ、一件当たり 12,000 に及ぶデータを機械学習による予測モデルを用いて独自のクレジットスコア（1～1000 で値が高いほど信用度が高い）を算出している。

以上のようなサービスが開発途上国において拡大している背景には、先進国では金融に関する厳しい法整備が確立されており、近年さらに強化されつつあるのに対して、開発途上国の多くではそれらが緩やか、あるいは確立されていないことがある。

1.1.3 新たな課題とモデルの解釈可能性

以上のように東南アジア諸国の金融課題は部分的ながらも解決されつつあるが、新たな課題も出現している。これまで与信審査関連業務に使用されてこなか

ったデジタル・フットプリントのようなデータを組み込むことで信用スコアを予測するモデル（信用スコアリングモデル）が複雑になり、どうしてある予測がなされたのかがわからなくなること、また予測を出力するプロセスが難解であることから、そのようなモデルの構築に使用する機械学習手法自体が安易に信用できないとする懸念の声がある。また、そもそもデジタル・フットプリントなどから得られたデータが信用スコアと結びつくような情報を持っているのかについてもわからないことから、その信用スコアを安易に信用することへ懸念が生じている。このような課題は金融分野に限った話ではなく、一般に「AI（機械学習）の解釈（説明）可能性」の問題として、近年様々な分野に適用され、議論がなされている。

1.2 研究目的

本研究の目的は、開発途上国における信用スコアリングモデルの抱える課題について機械学習における解釈可能性の知見を用いて分析・考察を行うことにより、新たなデータや手法を使用したモデルを本当に信用してよいのかという懸念を晴らすこと、および解釈可能性に関する研究において新たな知識を発見することにある。そのために、まず解釈可能性の背景や必要性などの議論を整理し、次にベトナム海運商業銀行から提供を受けたデータに対してクラスタリングと

相関ルール分析を用いて分析を行い、最後にそれらの総括を行う。

1.3 知識科学的意義

本研究の一つの特徴は、個人の信用スコアとソーシャルネットワークサービスの投稿内容を結びつけたデータを使用している点にある。私の知る限りでは、このようなデータを使用した研究はないので、ここから意味のある情報や知見を得られたならば、知識発見という意味で意義があると言える。また、AI（機械学習）の解釈可能性に関する研究は比較的新しいトピックであり、特に日本ではAI 関連のガイドラインなどに記述が増えているのにもかかわらず、その基本的な理論について記述した研究は、私の確認した限りでは見つからなかった。そこで本研究では、解釈可能性についての必要性や人間にとって良い説明とは何かなどの根底を成す議論を整理したうえで、それらを金融分野に適用して分析を試みた。したがって、AI の活用が広がりつつある社会において本研究は知識科学的意義を有するのではないかと思われる。

1.4 本論文の構成

本章では主に、途上国における金融課題、その解決が期待されているフィンテック、新たな課題と AI（機械学習）における解釈可能性について述べた。第 2 章では本研究に関連した研究を整理するが、特に機械学習における解釈可能性

について詳しく述べる。第 3 章では本研究の分析手法および使用するデータについて述べ、第 4 章ではその分析結果と考察を示す。最後に第 5 章では、本研究の全体の総括と今後の課題について述べる。

第2章 関連研究

2.1 解釈可能な AI(XAI)について

最初に、この章で使用する用語の定義について記述する[2]。「解釈（説明）可能AI」は、機械学習モデルの動作や予測を人間にも理解できるように構築されたモデルや方法、およびそれらを内包する人工知能技術一般を指している。また「XAI」はアメリカの国防高等研究計画局（DARPA）が主導している研究プロジェクト[3]において使用されている Explainable AI の略称である。「ブラックボックス」は、内部の仕組みを明らかにしないシステムのことであり、機械学習分野においては、パラメータを見ても理解できないモデル（ニューラルネットワーク等）を意味している。

2.1.1 社会情勢・背景[4]

人工知能技術において解釈（説明）可能性が問題とされるようになった背景には、近年の AI 技術の発展・普及、それらへの社会的な期待とともに生まれた、不安（ブラックボックスに陥る可能性のあるモデルを安易に信用してよいものなのか）がある。このような漠然とした不安に対して、日本では「AI 開発ガイ

[2] 用語については [7]および [9]を参考とした。
[3] https://www.darpa.mil/program/explainable-artificial-intelligence
[4] [6]および [7]を一部参考とした。

ドライン案」 [3]が 2017 年に総務省より策定され、透明性の原則、アカウンタビリティの原則が盛り込まれた。また、内閣府の統合イノベーション戦略推進会議（2019 年）において策定された「人間中心の AI 社会原則」 [4]においても公平性・説明責任及び透明性の原則が盛り込まれている。この二つのガイドラインと原則の一部を組み込んで 2019 年 8 月に総務省情報通信政策研究所から公表された「AI 利活用ガイドライン」 [5]でも同じく公平性、透明性、アカウンタビリティの三つが原則として盛り込まれている。現在の最も新しいガイドラインである「AI 利活用ガイドライン」で留意すべき項目は、「適正利用」の原則において「人間の判断の介在」が盛り込まれていることである。この項目は「人間中心の AI 社会原則」で言及されたことを引き継いでいるが、意思決定において AI と最終判断を行う人間がどのような関係を築き、どのようなプロセスを原則的に経るべきかがより明確に明記されている。

以上は日本国内の情勢であるが、海外でもある程度ではあるが同様の内容が検討されており、例えば EU では 2018 年より導入された General Data Protection Regulation[5](GDPR：EU 一般データ保護規則)に同様の内容がみられる。

[5] https://gdpr-info.eu/

2.1.2 研究界の動向

日本国内における解釈可能 AI を扱う研究者として有名なのは、原聡（大阪大学産業科学研究所）であり、人工知能学会誌の「私のブックマーク」2018 年（33 巻 4 号）に「機械学習における解釈性」 [6]というタイトルで、2019 年（34 巻 4 号）に「説明可能 AI」 [7]という論文が寄稿されている。これらの論文では、その時点での国内外の社会情勢と研究会の動向、有用な資料、提案されている説明法、そして課題について記述されている。

海外での研究はより盛んであり、様々な解釈法に関する論文が投稿されているだけでなく、解釈可能性を議論する上での論点を様々な側面から論説した文献や資料も公開されている。例えば AAAI 2019 Tutorial On Explainable AI: From Theory to Motivation, Applications and Limitations [8]では解釈可能な AI について多角的な解説、研究の紹介がなされている。また、Christoph Molnar(2019) Interpretable Machine Learning: A Guide for Making Black Box Models Explainable [9]は、解釈可能性の重要性や適用すべき範囲、説明法の具体例など、解釈可能性を機械学習に求める際に考慮すべき論点がよく整理されている優れた文献であり、本研究の 2.1 節はこの文献を参考にしている。

2.1.3 解釈可能性の必要性[6]

一般に機械学習モデルに解釈可能性を付与すると、その予測精度は低下する。このようなトレードオフの関係の中で解釈可能性を求める理由としては、大きく三つが挙げられる。

第一に、解釈可能性の必要性は問題の定式化が不完全であることから生じる。つまり、モデルは正しい予測を示していたとしても、それは問題を部分的にしか解決していない可能性があるがゆえに解釈可能性が必要とされるということである。これは本稿のテーマである与信審査（ローンの審査）を例にするとわかりやすい。特別な操作を加えない場合、機械学習モデルはトレーニングデータからバイアスを拾い上げる。したがって、構築された与信審査モデルは意図せずしてマイノリティを差別する可能性がある。例えば、与信審査モデルを構築する際に、前提となる目標は「最終的に返済する人にだけ融資を与えること」であるが、追加の制約条件として「特定の人口統計に基づいた差別をしない義務」が存在するという場合である。また、このような具体的な例にとどまらず、科学技術の発展という大きな枠組みで捉えると、「問題」は科学の目的である知識の獲得となる。多くの問題がビッグデータや機械学習モデルによって解決されることが考えられることから、データそのものや予測結果ではなく、モデルそのものに知識の核

[6] [9] を参考とした。

が含まれることも大いに考えられ、解釈可能性は追加の知識の獲得にも貢献が期待できる。

第二に、解釈可能性はAI関連技術やそれらを組み込んだシステムや製品が日常生活に導入される過程における社会的な受容性を高めると指摘されている。例えば、ロボット掃除機が停止している場合、それが何もない床で機能を停止しているよりも、台所のラグを巻き込んで停止している方が人間にとって受け入れやすいかもしれない。これは、待ち合わせで理由もなしに遅刻されるより何らかの理由（電車の遅れなど）があったほうが腹が立たないことと本質的に近い。ここで、解釈可能性の必要性を社会的な受容性の向上に求める場合に、気を付けなければならないことは、説明する側の機械と説明を受けて理解する側の人間の間にズレが生じる可能性があること、そしてズレがあっても大きな問題にはならないことである。同じくロボット掃除機で例えると、片方の車輪が正しく動作しないことや障害物の検知に関わるシステムのバグなど複合的な要因がロボット掃除機の動作の停止を引き起こしていたとしても、ロボット掃除機の行動を理解する側の人間は「何か邪魔なものがあると止まるのであろう」と解釈できるだけで十分なのである。現在では、この必要性は少々無理があるように感じられるかもしれないが、AI関連技術がさらに発展・普及した近い将来において、人間と機械が社会的な相互関係を形成する際には重要なポイントにもなりうる。

第三に、機械学習モデルをデバックや監査に使用する場合には解釈可能性が必要となる。機械学習モデルがある製品に搭載されたのちに、期待通りの性能を発揮していないということは大いに考えられる。この時、解釈可能性はエラーの原因を理解することに役に立ち、またシステムの修正の方向性も示す。これは、有名な「ハスキーと狼の分類器」が一部のハスキーを狼に誤分類していた話が分かりやすい。この例では、画像を狼として分類するために背景の雪を重要な特徴として学習していた。これは訓練データを適切に分類するという点では有意義だが、実際に使用するモデルとしては役に立たない。解釈可能性を備えたモデルであれば、この失敗に気が付くことも修正することも比較的容易となる。

2.1.4 解釈可能性を必要としない場合[7]

ここまで解釈可能性の必要性について記述してきたが、逆に解釈可能性を必要としないシナリオというものも存在するので、それらについても整理しておく。

第一に、モデルが大きな社会的あるいは経済的な影響を与えない場合は、解釈可能性は必要とされない。例えば、少し高級な機械学習手法を用いた性格診断がひっそりと無料で公開されたとする。この場合、モデルが間違っていたとしても

[7] [9] を参考とした。

予測結果について説明もなく解釈もできないとしても問題はない。しかし、この性格診断が思ったより性能が良いためにビジネスに活用されるとなると話は変わってくる。この診断結果が就職活動に利用されるといった話になれば、前述の「必要性」が求められる事態へと変わるかもしれない。

第二に、問題が十分に研究されている場合も解釈可能性は必要ない。機械学習モデルが使用されているアプリケーションやソフトウェアの中には十分に研究がなされ実用化もされているので、モデルが適用されている（解決している）問題の経験・知識が十分に蓄積されているものもある。よい例としては、郵便物における郵便番号の識別モデルが挙げられる。これは長年の実務経験もあり、動作することが明らかである。さらに、解釈可能性によって追加の知識を得ることも期待はできない。

第三に、モデルの作成者側の目的と利用者側の目的がミスマッチを起こしている場合には、解釈可能性は必ずしも不要であるとは言えないが注意しなければならない。例えば、与信審査のシステムであれば、融資側は返済する可能性が高い申込者にだけ融資したいと考えるが、利用者側は銀行が融資をしたくなくとも融資を受けたいと考える。ここで「3 枚以上のクレジットカードを保有していると信用スコアが下がる」という情報が流れたとすると、利用者側には 3 枚目のカードを一度解約して融資が承認された後にカードを再申請するといった

インセンティブが生まれる。この時、利用者の信用スコアは上昇するかもしれないが実際に返済する確率は変わらない。このような実際に結果と因果関係がない特徴がモデルに影響を与えているがゆえにゲーム[8]が成立してしまうケースにおいては、少なくとも解釈可能性の取り扱いに注意を払う必要がある。

2.1.5 解釈手法の分類

モデルに解釈可能性を備えさせる手法は、様々な基準によって分類されるが、ここでは二つの基準を記述する[9]。

一つ目の基準では、本質的(Intrinsic)な手法とポストホック(post hoc)な手法の二つに分類される。前者は機械学習モデルの複雑さを制限することで解釈可能性を得られる手法であり、単純な決定木や線形回帰などの単純な構造であるため、過度に複雑にしなければ解釈可能だと判断されるような手法が該当する。後者のポストホックは医療系の文献でしばしば登場する事後解析(post hoc analysis)と同様の意味であり、モデルの学習後に解釈可能性を付与する手法が該当する。したがってポストホックな手法は本質的な手法にも適用できる。

次に、二つ目の分類を示す。

---

8 経済学におけるゲーム理論の意味。

9 [6]および [9]を参考とした。

・特徴量の要約統計量の算出および可視化

各特徴量の要約統計量を提供またはそれらを可視化することにより、解釈可能性を持たせる手法。最もシンプルな手法と言えるが、可視化することで意味を成す要約統計量も存在することから様々な可視化手法が提案されている。

・本質的に解釈可能性による近似

ブラックボックスなモデルを解釈する一つの方法は、本質的な解釈可能性を備えたモデルでブラックボックスを（大局的または局所的に）近似することである。そして、解釈可能なモデル自体はパラメータや要約統計量などから解釈できる。

・局所的な説明の提供

ある入力 x をモデルが y と予測したとき、その予測の根拠をモデルに提示させることで解釈可能性を実現する手法。前述のハスキーと狼の分類における失敗を指摘した手法 LIME はこの手法に分類される。

・解釈可能なモデルの設計

最初から決定木や線形回帰といった本質的に解釈可能なモデルを設計するというアプローチから生まれた手法。

2.1.6 人間にとって良い説明とは[10]

この項目では人間にとって「良い」説明とは何かについて整理し、それは解釈可能な機械学習にどのような意味をもたらすのかを考えていく。なお以下の記述は、機械学習モデルのデバックや航空機事故の原因の究明のような、ある結果に対して完全な因果関係の帰属が求められ、完全に（近い）論理的な説明・解釈が求められるようなケースとは関係が薄い。しかしながら、我が国のようにガイドラインの様々な文脈で「説明・解釈可能性」が現れ、説明のもつ社会的側面も考慮すべきときには、確かな知見を与えてくれる論点でもある。

・対照性

人間は一般に「どうして客観的あるいは主観的に望ましい結果ではなく、目の前の結果があるのか」と考え、さらに「もしあの時こうしていれば、望む結果が生じたのではないか」といった反事実的(counterfactual)な思考を巡らす。このような思考に則った「良い説明」の一つは、自分が疑問に感じる事例と、この事例と状況的に似ているが違う結果をもつ事例を比較することである。例えば、医師が「どうしてある患者には薬が効かないのか」を考え、説明したいと思うのなら、その患者と状況などは似ているが薬の効いた患者とを比較して説明を組み

10 主に[9]を参考とした。

立てるといったことである。このような対照的な説明は完全な説明ではないが、理解しやすい。つまり良い説明は、関心のある対象と参照対象との間にある最大の違いを強調しているものなのである。

反事実的な思考は現在から過去に向かう場面で顕著であるが、現在から未来に向かう予測でも時系列の線上に存在するという点で本質的に変わらない。そのため、解釈可能な機械学習においても、対照的あるいは反事実的な説明は有効であると考えられており、複数の研究が発表されている。

・選択性

ある出来事が唯一の原因に基づいていることは稀で、むしろ複数の原因が重なりあった結果であることが多い。しかしながら、出来事の原因として語られるのはその中の一つか二つの妥当と思われるものであり、私たちもこの「説明」に慣れているのだが、この妥当と思われるものがすべての人の間で一致するとは限らない。一つの出来事を様々な人々が様々な原因で説明することで矛盾が生じる現象は「羅生門効果」と呼ばれ、由来は当然ながら小説および映画の「羅生門」にある。機械学習モデルにおいて異なる特徴量から正確な予測が得られることは頑健性にも繋がり、有利であるが、同時に予測がなされた理由も複数存在することを意味し、それらは相互に矛盾する可能性がある。

・社会性

前述の通り、説明する側と説明を受ける側の間には相互作用的な関係性が存在するため、説明には社会的な側面が少なからず求められる。学会における専門用語満載の説明が、専門知識を持たない人々にとって相応しいものとは限らない。解釈可能な機械学習においても、この点については注意を払う必要がある。

・異常な点への焦点

人は出来事を説明するために原因をリストアップしたとき、普段とは異なる挙動をしているもの、すなわち異常なものや事態に注目する。異常について別の言い方をすると「起こる確率は小さいにもかかわらず起こったこと」と言える。異常な状態にあったものが通常通りであったならば重大インシデントに繋がっていないだろう、といったカウンターファクチュアルな考えをするからこそ、異常な原因あるいは異常な事態に繋がった原因は人間にとって良い説明になりうる。

・真実性

説明は現実において真実を写していることが望ましい、つまり、条件が同じ別の状況でも等しく原因と結果を結び付けているべきである。しかしながら、真実であることは人間にとって良い説明において最も重要な点ではない。「選択性」

においても触れたが、一般的な説明は、考えられる原因のすべてを網羅しているわけではなく、その中から一つか二つを選択している。そして、その選択は真実の一部を確実に省略している。株式市場の暴落は多数の原因が多数の人々の行動に影響を与えた結果だとするのが真実に近いと言えるが、ニュースで語られる原因はせいぜい数個でしかない。

・確証バイアスの存在

人間は信念と矛盾する情報を無視する傾向にあり、これは確証バイアスと呼ばれているが、説明もこのようなバイアスからは逃れられない。信念と一致していることは人間にとって良い説明の一つの構成要素になりそうだが、機械学習モデルに組み込むことは難しく、もし組み込んだとしても相当の性能の低下を引き起こすと考えられる。

・一般性

多くの出来事（一般的で確率の高い事象）を説明できる原因は一般的であり、良い説明に繋がると考えられる。しかし、これは「異常への焦点が良い説明となる」ことと矛盾する。異常な事態は起こるのが稀であるからこそ「異常」なのであって、異常な出来事がない場合は一般的な説明が良い説明となると考えられる。ただ、人は何も問題のない通常の状態で説明を求めたくなることは稀で、多

くの場合は何か問題のある異常な状態に対して説明を求めたくなるのである。

2.1.7 現状の課題[11]

最後に、解釈可能性を持つ機械学習に関する研究およびその周辺が抱える課題について記述する。

第一に、現時点での解釈（説明）可能性の研究の多くは、研究者各自の仮説、つまり「このような解釈・説明ができれば便利であろう」というものを土台としているために、実用性に乏しい。様々なガイドラインに解釈可能性が記述されていることに鑑みても、具体的な問題を抱える産業界からの研究への参入が望まれている。

第二に、解釈可能性に対して過度に信頼あるいは期待を寄せることについても注意が必要である。特に深層学習モデルの説明においては意図的に間違った説明を生成することも可能であると報告されており、誤った説明がなされるリスクを考慮して実用化以前に適切な検証が必要とされる。また、解釈可能性は一般に精度とのトレードオフの関係にあることはすでに述べたが、それだけはなく上記の誤った説明のリスクや計算コストがあるため、やはり最終的には人間による判断が必要とされる。したがって、モデルに解釈可能性を求める際には本

[11] [6] を参考とした。

当にそれが必要なのか、何のために必要なのか、導入がコストに見合うか、など

について検討する必要がある。

第3章　分析手法とデータ

3.1 信用スコアリングモデルにおける解釈可能性

本章では、本研究で使用するデータおよび分析手法について概要を述べるが、その前に解釈可能性の議論を踏まえたうえでの分析の方針を整理する。

まず、解釈可能性の必要性をどこに求めるかであるが、その前提として「最終的に返済する人にだけ融資を与えること」と「特定の人口統計に基づいた差別をしないこと」の 2 つが考えられる。今回、分析に使用するデータセットでは前者はともかく後者は困難である。したがって、社会的な受容性を高めること、つまり新たな手法やデータを与信用スコアリングモデルに利用することへの不信感を軽減することに分析の焦点を当てる。

このような不信感を軽減するためのアプローチは大きく 2 つ存在し、1 つは精度の側面からで S. Lessmann et al. (2015)や澤木ほか（2017）のアプローチである。この 2 つの関連研究は共にベンチマークという客観的な数字から、既存手法にとって代わるべき新規手法の優位性を示し、またデータセットの選択から外部妥当性[12]も限定的であるが示していた。本研究で使用するデータセットはサンプルサイズが小さいので、実際にモデルを構築し精度を測ることの妥当

[12] ただ実験室の中（使用したデータセットにおいて）でのみ有効であるのではなく、現実世界においても先進的な手法が通用すること。

性は薄い。しかしながら、顧客の SNS への投稿データと信用スコアが紐づいている珍しいデータであるので、もう一つのアプローチであるデータセットから何らかの意味のある情報を取り出すことで新たなデータを与信審査モデルに組み込むことへの不信感を軽減することを目指す。これは社会的な受容性の項で述べた通り、必ずしも真実の全てを写してはいないが、だからこそ解釈可能性の向上に寄与するのではないかと思われる。

最後に、誰に対して解釈可能性を示すのかについてだが、これは解釈可能性を必要としない場合の項のゲームが成立し、モデルを人が偽るインセンティブが生まれる可能性が考えられるため、本研究では資金提供を受ける側ではなく資金提供をする側と設定する。

3.2 使用するデータ

使用するデータはベトナム海運商業銀行(Vietnam Maritime Commercial Joint Stock Bank)から提供を受けたもので、顧客の銀行口座情報と SNS(Facebook)上の公開情報を照合したものである。なお、情報の照合は、顧客が銀行に申告した Facebook の ID を用い、銀行が顧客の公開情報を利用することに同意した顧客に対して行われている。

ここで使用するデータは、顧客の Facebook の ID、性別、年齢、家族構成、

信用スコア、審査申請受理地域、および SNS への投稿内容である。家族構成については「不明」の顧客が全体の 3 割を占める。信用スコアについては A、B、C の 3 値（A が最も高く、C が最も低い）なので以下、信用度と言い換える。また信用度 B の顧客数が全体に占める割合は 7.8%なので分析・考察には信用度 A と C を用いる。審査申請受理地域についてはハノイ市が全体の 22.4%、ホーチミン市が 29%を占めていたので、その他の都市、省については「その他地域」としてまとめた。

SNS の投稿内容のデータの例を表 3.1 に示した。左端が顧客番号（Facebook の ID を簡略化したもの）で、その右に連なる一行がある顧客の一回の投稿の内容を示している。書き込みの内容は全てデータセットを作成した銀行によって分類がされているため、例えば顧客番号 36 の顧客は銀行側に「英語の勉強」とみなされる投稿を一回、「犬を飼う」という投稿を一回したということである。また、投稿の内容の分類は LV5 から LV1 にかけて丸め込みがなされる。例えば、顧客番号 59 の人の投稿は LV5 では「ヤマハオートバイ」だとみなされ、LV4 ではより広い意味のオートバイ会社だとみなされ、LV3 ではさらに広い意味の「オートバイ」、LV2 では「車両」とみなされていき、最後の LV 1 では「買い物」だとみなされる。本稿では「オートバイ」や「買い物」などの投稿の内容を示すものを以下、属性と記述する。LV1 の列は最も大括りの分類であり、すべての投稿

内容は 11 の属性（サービス、不動産、健康、学校・教育、生活、職業、芸術・娯楽、買い物、購買傾向、社会、飲食）に分類される。なお、今回の分析の対象としているのは、データセットのもつ 681 名の顧客のうち Facebook の ID が存在する 562 名である。

顧客番号	LV1_NAME	LV2_NAME	LV3_NAME	LV4_NAME	LV5_NAME
36	学校・教育	他の教育	外国語勉強	英語の勉強	英語の勉強
36	生活	日常活動	ペット	犬を飼う	犬を飼う
59	買い物	車両	オートバイ	オートバイ会社	ヤマハオートバイ
…	…	…	…	…	…

表 3.1 SNS への投稿内容のデータの例

3.3 分析手法について

3.3.1 クラスタリング

クラスタリングは教師なし学習の一つで、外的な基準ではなくデータ点の類似度に基づいてデータ点を複数の集合（クラスター）に分割する手法であり、経済や経営などの広い分野で使用されているものである。本研究では、このクラスタリングを SNS への投稿の傾向に基づく顧客の分類に使用する。

・分析手順[13]

クラスタリングに関しては LV1 に含まれる 11 の属性に分類される投稿の回数に基づいて行っており、具体的な手順は次の通りである。まず、それぞれの投稿の回数を連続値として扱っているのだが、ある属性については投稿がない場合やある属性については極端に多くの投稿を行っている場合などがあるので、すべての顧客のすべての属性について投稿の回数に 1 を足したうえで対数化を行う。次に、11 の属性の特徴を主成分分析（PCA）によって、より少ない合成変数（主成分）に要約し、顧客ごとの主成分得点を求める。今回扱うデータは、属性によっては正の相関のある変数があったので、出来るだけ少ない情報損失で互いに相関のない形で合成できる主成分分析を行った。さらに、主成分分析によって得られた主成分得点から、累積寄与率が基準とされる 70%を上回る第 1 から第 4 主成分までの主成分得点を使用してクラスタリングを行う。クラスタリングについては最も一般的な手法の一つである k-means を使用した。主成分分析およびクラスタリング（k-means）の実行は Python[14]および機械学習ライブラリの scikit-learn[15]を利用した。

[13] 主に [12]を参考とした。

[14] https://www.python.org/

[15] https://scikit-learn.org/stable/

3.3.2 相関ルール分析[16]

相関ルール分析は、データの中から、同時性や関係性をルール（規則性）として抽出する手法であり、相関ルールは（A→B）の形式で表現される。一般的にA を条件部、B を帰結部と呼び、A と B は共に複数のアイテムを含んでよい。相関ルール分析にはルールの重要性を示す評価指標があるので、一般的なものを以下に示す。

・支持度(support)

$$Support(A \rightarrow B) = P(A \cap B)$$

・信頼度(Confidence)

$$Confidence(A \rightarrow B) = \frac{P(A \cap B)}{P(A)}$$

・リフト値(Lift)

$$Lift(A \rightarrow B) = \frac{Confidence(A \rightarrow B)}{Support(B)} = \frac{P(A \cap B)}{P(A)P(B)}$$

各指標について簡単に説明すると、支持度はデータセットの中でアイテム（セット）が出現する割合を表し、信頼度はルール（A→B）の条件 A と B の両方を満たすデータ点の割合を表している。したがって、ルール（A→B）と（B→

[16] [13]、 [14]、 [15]主にを参考とした。

A）とでは異なり向きが存在する。最後のリフト値はルール（A→B）の条件 A と条件 B が統計的に独立している場合より、どれくらい高い頻度で同時に表れているかを示している。したがってリフト値が 1 より大きければ、条件 B を満たす確率が条件 A を満たすことによって上昇していることを示している。しかしながら、これらの指標の値がいくらであれば重要なルールであるとされるかについての明確な基準は存在しない。

購買データを分析している既存研究ではデータ全体ではなく、期間や性別などで対象を限定して相関ルール分析を行い、ルールの差異などを視察することが提案されている [13]が、本研究ではクラスタリングで得られたクラスターそれぞれに対して分析を行う。また、相関ルール分析は顧客の SNS の投稿に関するデータの LV4 を使用して行うので、一般的な特徴的なルールを取り出す目的のほかに、可視化のためのツールとしても相関ルール分析を利用する。なお、相関ルール分析の実行は Python および機械学習ライブラリの mlxtend[17]を利用した。

[17] http://rasbt.github.io/mlxtend/

第4章 分析結果と考察

4.1 クラスタリングの結果

・主成分分析について

クラスタリングの結果を示す前に、前処理として行った主成分分析（PCA）の結果について簡単に示す。表 4.1 はクラスタリングで使用した第 1～4 主成分の固有ベクトルである。詳細は省くが、各顧客の主成分得点は固有ベクトル（サービス）×各顧客の値（サービス）＋固有ベクトル（不動産）×各顧客の値（不動産）＋・・・＋固有ベクトル（飲食）×各顧客の値（飲食）と求められるため、固有ベクトルの値がプラスのカテゴリは主成分得点が高くなり、逆にマイナスなら低くなる。第 1 主成分はすべてのカテゴリの値が 0.3 前後となっており、単純に SNS をよく利用しているか否かという点にデータの分散が大きいことを示している。第 2 主成分は、社会がプラスに大きく、また学校・教育と飲食がマイナスに大きく振れているので、社会に関する書き込みの数と学校・教育と飲食のそれが主成分得点の値の増減をもたらしており、またデータの分散が大きいことがわかる。第 3、第 4 主成分も同様に、それぞれ不動産と職業に関する書き込みをしているか否かという点にデータの分散があることがわかる。

	サービス	不動産	健康	学校・教育	生活	職業	芸術・娯楽	買い物	購買傾向	社会	飲食
第1主成分	0.3669	0.2491	0.2776	0.2692	0.3295	0.2518	0.3245	0.3684	0.3127	0.2230	0.3054
第2主成分	-0.0439	0.2810	-0.1406	-0.3195	0.1581	-0.1520	-0.1294	-0.0958	0.2052	0.7592	-0.3236
第3主成分	-0.0671	0.7872	0.1102	0.0244	-0.2058	0.0255	-0.4202	-0.0755	0.1848	-0.3087	0.0915
第4主成分	-0.1289	0.0723	-0.0789	0.2946	0.0471	0.8072	-0.0638	-0.2208	-0.2692	0.1345	-0.2967

表 4.1 各主成分の固有ベクトル

・クラスタリングの結果

次に、各顧客の主成分得点を使用してクラスタリングを行った結果を示す。クラスタリングにおいていくつのクラスター（集合）に分割するかは、人間があらかじめ設定する値（ハイパーパラメータ）なので、今回は 4、6、8 の三通りの設定で行った。表 4.2 は 6 に設定した結果を示したもので、概ね全ての属性の値が高いか低いかという基準で分割されていることがわかる。この傾向はクラスターの数を 4 に設定した場合も同様だった。したがって、比較的 11 の属性ごとの傾向が表れていたクラスターの数を 8 とした結果を主に示していく。

クラスター	顧客数	サービス	不動産	健康	学校・教育	生活	職業	芸術・娯楽	買い物	購買傾向	社会	飲食
0	83	12.217	0.976	4.169	6.386	8.012	5.470	18.831	17.578	3.723	4.807	3.361
1	113	4.566	0.451	1.726	1.522	3.540	1.097	9.841	8.150	1.593	2.053	0.912
2	95	11.737	1.579	4.084	3.021	7.284	0.811	16.568	17.211	5.347	7.011	3.505
3	64	0.969	0.625	0.484	0.625	0.953	0.578	3.328	2.547	0.500	0.563	0.313
4	85	10.235	1.518	4.482	3.871	5.271	1.141	14.000	17.118	4.200	0.506	5.212
5	122	22.295	5.221	9.074	8.451	11.410	6.213	24.639	29.992	8.443	6.549	7.467
全体	562	11.205	1.936	4.356	4.251	6.509	2.753	15.406	16.534	4.299	3.870	3.717

＊クラスターと顧客数を除いて、各属性への投稿回数の平均

表 4.2 分割数を 6 としたクラスタリング

表 4.3 はクラスターの数を 8 に設定した結果を示したものである。各クラスターの「サービス」～「飲食」の 11 属性の値は、対数化などの操作を行ってい

ない値の平均であるので、例えば、サービスの値が高ければ、そのクラスターに所属する顧客はサービスに分類される投稿を多く行っていると考えられる。したがって、まず特徴的なのがクラスター2 は全ての属性の値が極端に低く、逆にクラスター7 は高いため、クラスター2 はアカウントを持っているだけでほとんど SNS を利用していない集団、対照的にクラスター7 は非常によく利用している集団であると捉えられる。次に、クラスター3 とクラスター4 は全体的にはよく似た値を示しているが、クラスター3 は不動産と対人関係の値が高いのに対して、クラスター4 は学校・教育の値が高いといった特徴がある。同様に、クラスター0 とクラスター5 も全体的によく似た値を示しているが、対人関係の値が大きく異なっている。クラスター1 はクラスター0 と 5 と比べると全体的に低い値が多いが、職業の値が高く、また学校・教育の値も若干高いことがわかる。クラスター6 はクラスター2 ほどではないが全体的な値が低く、やはり主成分分析で示されていた通りに SNS をよく利用しているかという方向によって分類が行われたことがわかる。

クラスター	顧客数	サービス	不動産	健康	学校・教育	生活	職業	芸術・娯楽	買い物	購買傾向	社会	飲食
0	94	11.085	1.000	4.096	2.936	7.000	0.894	17.043	16.894	4.915	6.872	3.213
1	57	7.053	0.474	2.719	3.825	4.842	4.386	14.561	11.439	1.649	2.351	1.807
2	58	0.672	0.500	0.345	0.690	0.810	0.569	3.190	2.397	0.466	0.414	0.345
3	92	16.196	4.891	5.576	5.239	9.739	4.630	17.380	21.489	6.848	7.804	4.196
4	64	17.047	1.891	6.734	8.063	8.297	5.156	22.016	23.109	5.328	2.016	6.563
5	71	9.690	1.577	4.014	3.493	5.408	0.563	13.268	16.549	4.324	0.507	5.310
6	86	4.000	0.605	1.302	1.128	3.093	0.453	8.244	6.907	1.616	2.116	0.558
7	40	30.025	5.075	13.675	12.800	15.000	8.625	34.550	42.200	10.400	7.650	10.825
全体	562	11.205	1.936	4.356	4.251	6.509	2.753	15.406	16.534	4.299	3.870	3.717

＊クラスターと顧客数を除いて、各属性への投稿回数の平均

表 4.3 分割数を 8 としたクラスタリングの結果 1

次に、SNS への書き込み以外のデータ、つまり性別、年齢、家族構成、申請受理地域、そして信用スコアと照らし合わせた結果を表 4.4 に示す。まず、信用スコアについて見ると、信用度 C の割合が比較的高いのはクラスター2 およびクラスター7 であり、それぞれ極端に SNS を利用していない集団と逆に極端によく利用している集団であった。逆に信用度の割合が比較的高いのはクラスター4 およびクラスター5 で、それぞれ学校・教育の値が高い、社会の値が低いといった特徴を持っていた。また、信用度 C の割合がクラスター2 に次いで高いクラスター6 は全属性の値が小さい、つまりクラスター 2 ほどではないが SNS をあまり利用していない集団であった。次に年齢について見ると、平均年齢の低いのはクラスター7 および 4 なのに対して、高いのはクラスター6 および 2 となっており、年齢が低い方が SNS をよく利用している傾向を示していると言え、一般的な認識とも一致している。申請受理地域については、ホーチミン市およびハノイ市の割合が高いクラスターは 3、4、5 であり、最も低いのはクラスター

2 であることから、都市部の方が SNS をよく利用していることを示していると考えられるが、今回はデータの関係からホーチミン市、ハノイ市、その他地域の 3 つにまとめたことを踏まえると、都市部か否かという基準に単純に結びつけることはできない。また、家族構成についても「不明」となっている顧客が約 3 割あるため、有意義な解釈は難しい。

クラスター	性別(男性割合)	年齢	家族構成				信用度			申請受理地域		
			離婚	既婚	独身	不明	A	B	C	ハノイ市	ホーチミン市	その他地域
0	42.6	26.87	2.1	28.7	34.0	35.1	50.0	4.3	45.7	22.3	23.4	54.3
1	28.1	27.25	1.8	29.8	52.6	15.8	56.1	12.3	31.6	22.8	22.8	54.4
2	29.3	30.07	5.2	29.3	27.6	37.9	44.8	6.9	48.3	13.8	29.3	56.9
3	51.1	27.51	5.4	22.8	37.0	34.8	55.4	16.3	28.3	23.9	30.4	45.7
4	46.9	26.28	3.1	15.6	45.3	35.9	68.8	6.3	25.0	28.1	32.8	39.1
5	59.2	27.63	1.4	29.6	43.7	25.4	59.2	9.9	31.0	22.5	32.4	45.1
6	47.7	30.86	2.3	37.2	26.7	33.7	48.8	3.5	47.7	16.3	30.2	53.5
7	57.5	25.95	2.5	17.5	40.0	40.0	47.5	0.0	52.5	35.0	32.5	32.5
全体	45.6	27.92	3.0	27.0	37.5	32.4	53.9	7.8	38.3	22.4	29.0	48.6

＊年齢のみ平均値で他は割合（%）

表 4.4 分割数を 8 としたクラスタリングの結果 2

これらの分析結果から、クラスタリングによって SNS への投稿に関して傾向の異なった集団を取り出すことができ、またそれらはある程度ではあるが顧客の信用度や年齢などと結びついていることがわかった。次節では、クラスタリングによって分類したクラスターに対して相関ルール分析を用いてより具体的なルールを取り出していく。

4.2 相関ルール分析の結果

この節では、クラスタリングで得られた各クラスターに対して相関ルール分

析を行った結果を示していく。

相関ルール分析もクラスタリングと同様に人の手によって設定する値（ハイパーパラメータ）があり、それが分析手法で扱った支持度（support）の下限をどうするかの設定、すなわち最低支持度（minsup）である。一般に最低支持度を小さくすればするほど膨大なルールが抽出され、考察するのが実質的に困難になるのだが [14]、今回は概ね SNS をよく利用しているか否かという基準を反映したクラスターそれぞれに相関ルール分析を行うため、この性質がクラスターによってはさらに極端になっている。したがって、やむなくクラスターごとに最低支持度を変えて分析を行った。また、全顧客 562 名のうち半数の 281 名が投稿したとみなされている 29 個の属性は分析環境の制限から共通に除外した。

各クラスターに適用した最低支持度と、その支持度の下で抽出したルールの数を表 4.5 にまとめた。表 4.5 にある通り、極端に SNS をよく利用していた集団であるクラスター 7 は最低支持度を 0.6 にまで上げても 167,737 個のルールが抽出され、逆のクラスター2 は最低支持度を 0.08 にまで下げても 3 個のルールしか抽出されなかったので、この 2 つのクラスターに関しては相関ルール分析を行うよりも「極端に SNS をよく利用している、あるいはしていない集団」と捉えるほうが賢明であると判断し、クラスター7 と 2 は以下の分析から除いた。

以下スペースの都合から、信用度 A の割合が比較的に高かったクラスター 4 と 5、C の割合が比較的に高かったクラスター 6 に関して、それぞれクラスターの持つ性質をよく表していると思われるルールをいくつか抜粋して示し、結果と比較・考察を記述していく。クラスター 4 、5 とクラスター 6 の最低支持度が異なるため比較について不安が残る点は、クラスター 0 の結果を補助的に用いる。なお、本分析では扱うデータの性質から信頼度（conf）が 1 となる無意味なルールが多く抽出されたため、考察を少しでも容易にするためにも一律に信頼度の上限を 0.85 に設定した。

クラスター	0	1	2	3	4	5	6	7
最低支持度	0.4	0.2	0.08	0.4	0.4	0.2	0.12	0.6
ルール数	132	38	3	1442	806	345	52	167737

表 4.5 各クラスターへ設定した最低支持度と得られたルールの数

・クラスター 4

クラスター4 では最低支持度 0.4 の設定の下で 806 個のルールが抽出された。それらのルールの一部を表 4.6 に示す。まず、特徴的なのは条件部、帰結部ともに（外国語勉強）や（その他教育）などの教育関連を含むルール、そして（健康トレーニング）を含むルールが多数抽出されたことである。また、他のクラスターと比較すると（英才教育→料理を学ぶ）が上位にあること、（投資）が（外国為替投資）と結びついていること、ベトナム料理を含むルールが抽出されている

ことなどが特徴的であった。

条件部	帰結部	support	conf	lift
英才教育	料理を学ぶ	0.406	0.813	2.000
英才教育	料理を学ぶ・料理	0.406	0.813	2.000
ファッションアクセサリー・投資	金・外国為替投資	0.406	0.813	2.000
お菓子	コーヒー/ソフトドリンク・ケーキ	0.422	0.818	1.939
ファッションアクセサリー・投資	宝石	0.422	0.844	1.862
舞台芸術	映画・英語の勉強	0.406	0.667	1.641
映画	その他教育・外国語勉強・舞台芸術	0.438	0.718	1.641
外国語勉強・職業	その他教育・飲食品	0.406	0.743	1.585
英語の勉強	その他教育・外国語勉強・投資	0.438	0.737	1.572
英語の勉強	その他教育・外国語勉強・映画	0.406	0.684	1.564
その他教育	健康トレーニング・外国語勉強	0.438	0.651	1.488
その他教育	外国語勉強・ベトナム料理	0.406	0.605	1.488
その他教育・コーヒー・ソフトドリンク	英語の勉強	0.406	0.839	1.413
英才教育	その他教育・外国語勉強	0.406	0.813	1.238
ベトナム料理	投資・職業	0.406	0.684	1.183
ベトナム料理	ファッションアクセサリー・職業	0.406	0.684	1.152
飲食物・職業	ベトナム料理	0.406	0.684	1.152
健康トレーニング	ゲーム・職業	0.422	0.600	1.129
ベトナム料理	舞台芸術	0.406	0.684	1.123
健康トレーニング	電子製品・職業	0.438	0.622	1.106

表 4.6 クラスター4 の結果

・クラスター5

クラスター 5 では最低支持度 0.2 の設定の下で 345 個のルールが抽出された。同様に結果を表 4.7 に示す。クラスター5 はクラスター4 に次いで信用度 A の割合が高い（68.8%に対して 59.2%）集団であったため、教育関連を含むルールがある点や（英才教育）を含むルールがある点などクラスター 4 の特徴を一部持っている。異なるのは（健康トレーニング）が（フィットネス・運動）のような同一属性としか結びついていない点や最低支持度を引き下げたことを加味しても飲食関連のルールの占める割合が上昇した点などである。

条件部	帰結部	support	conf	lift
映画	レストランサービス・舞台芸術	0.211	0.750	3.550
その他教育・外国語勉強	英語の勉強	0.211	0.714	3.381
英才教育・料理	料理を学ぶ・輸入品購入	0.225	0.727	3.227
ケーキ	お菓子・ベトナム料理	0.211	0.652	3.087
その他教育	外国語勉強	0.296	0.840	2.840
西洋料理	洋風のケーキ	0.254	0.720	2.840
その他教育	英語の勉強	0.211	0.600	2.840
健康トレーニング	フィットネス・運動	0.225	0.516	2.290
レストランサービス・ベトナム料理	おやつ	0.239	0.708	1.934
コーヒー・ソフトドリンク	ミルクティー	0.268	0.442	1.651

表 4.7 クラスター5 の結果

・クラスター 6

クラスター6 では最低支持度 0.12 の設定の下で 52 個のルールが抽出された。同様に結果を表 4.8 に示す。クラスター6 で特徴的なのは（ギャンブル→不健康な娯楽）というルールが抽出された一方で、クラスター4 や 5 で見られた教育関連などのルールが一切抽出されず、表にある（メンズファッション→メンズ洋服）以外は全て社会や人間関係のルールで占められていた点である。

条件部	帰結部	support	conf	lift
ギャンブル	不健康な娯楽	0.128	0.786	6.143
メンズファッション	メンズ洋服	0.140	0.706	5.059
人間関係	友達・社会とコミュニティ	0.140	0.667	4.778
人間関係・社会	デート	0.174	0.833	4.216
母親・乳児の商品	社会	0.140	0.600	1.911

表 4.8 クラスター6 の結果

・クラスター0

上記のクラスター6 からはクラスター4 と 5 からは抽出されない（ギャンブル）を含むルールが抽出されたが、設定した最低支持度が 0.12 と低い。そこで、

信用度 C の割合が 45.7%とクラスター6（47.7%）に次いで高いクラスター0 に対して最低支持度を変化させて分析を試みた。結果は最低支持度 0.4 では社会関係関連のルールしか抽出されなかったが、これを 0.3 に下げると（ギャンブル）を含むルールが支持度 0.351、信頼度 0.85、リフト値 2.28 で抽出された。したがって、クラスター6 に偶然にも（ギャンブル）の投稿をした人が紛れ込んだということではなく、信用度が低いことと（ギャンブル）の投稿をしたことには何らかの結びつきがあるのではないかと思われる。しかしながら、クラスター0 は最低支持度 0.4 の時点でクラスター5 や 4 にて見られた教育関連のルールが出現しなかったことも事実であるため、ここでの結果はあくまで傾向であることに留意すべきである。

・信用度 A および C のみ

最後に、クラスターごとに行った相関ルール分析がどれだけ信用度に関するルールを抽出できているかを見るために、信用度 A および信用度 C のみの顧客からなるデータに対して相関ルール分析を行う。この分析では、事前の結果からクラスター7 のデータからは膨大なルールが抽出されることがわかっていたため、A と C ともにクラスター7 に属する顧客は除いた。この操作によって属する顧客の数は A で 303→284、C で 215 から 194 に減ったが、抽出されるルー

ルはそれぞれ特徴を維持しながら信用度 A で 1104→308、信用度 C で 983→198 に削減することができた。表 4.9（信用度 A）および表 4.10（信用度 C）は、この操作の下の結果の抜粋であり、最低支持度は共に 0.2 である。

表 4.9 の信用度 A のみの顧客の結果からは、クラスター4 と 5 で多く見られた教育関連や（健康トレーニング）を含むルールが多数抽出されたほか、（英才教育→料理を学ぶ）といった特徴的なものも同じくあった。対して表 4.10 が示す C のみの顧客からなる結果からは、以上のようなルールは抽出されず、代わりに（ギャンブル→不健康な娯楽）というクラスター6 および 0 と共通するルールが抽出され、また社会関連のルールが多数を占める点もクラスター6 および 0 と似通っていた。また、（投資）を含むルールに関しても信用度 A は（金/外国為替投資）や（宝石）などの一般的に所得の高い人が関心を示す属性と結びついているのに対して、信用度 C は（社会関係）や（ファッションアクセサリー）などとしか結びついていない点などが異なっていた。

条件部	帰結部	support	conf	lift
英才教育	料理を学ぶ	0.236	0.817	3.463
英才教育	料理を学ぶ・料理	0.236	0.817	3.463
ファッションアクセサリー・投資	宝石	0.218	0.827	3.453
外国語勉強・その他教育	英語の勉強	0.303	0.811	2.679
その他教育	外国語勉強・輸入品購入	0.204	0.518	2.493
電子製品	スマートフォン	0.254	0.600	2.367
社会とコミュニティ	社会・輸入品購入	0.215	0.598	2.327
健康トレーニング	運動	0.254	0.581	2.290
投資	金/外国為替投資	0.201	0.460	2.290
健康トレーニング	フィットネス・運動	0.250	0.573	2.290
社会	電子製品	0.264	0.605	2.290
投資	宝石	0.218	0.500	2.088
レストランサービス	ベトナム料理	0.229	0.570	1.499
ペット	投資	0.208	0.615	1.408
コーヒー・ソフトドリンク	健康トレーニング	0.246	0.593	1.359
投資	ベトナム料理	0.225	0.516	1.357
コーヒー・ソフトドリンク	電子製品	0.236	0.568	1.344
ブランド品志向の購入	健康トレーニング	0.225	0.552	1.264
ベトナム料理	インテリアグッズ	0.211	0.556	1.262
電子製品	健康トレーニング	0.232	0.550	1.260
健康トレーニング	職業	0.225	0.516	1.253
外国語勉強	飲食物	0.201	0.533	1.250
輸入品購入	ペット	0.201	0.410	1.213
ベトナム料理	社会	0.201	0.528	1.209

表 4.9 信用度 A のみの結果

条件部	帰結部	support	conf	lift
ギャンブル	不健康な娯楽	0.278	0.806	2.843
ファッションアクセサリー	宝石	0.211	0.577	2.732
社会とコミュニティ	ギャンブル・社会	0.206	0.513	2.427
電子製品	パソコン	0.216	0.525	2.122
社会	社会とコミュニティ・ギャンブル	0.206	0.435	2.109
社会	社会とコミュニティ・デート	0.273	0.576	2.109
社会とコミュニティ	社会・輸入品購入	0.211	0.526	2.217
社会とコミュニティ	レストランサービス・社会関係	0.211	0.526	2.170
電子製品	オートバイ	0.206	0.500	1.565
ギャンブル	電子製品	0.211	0.612	1.484
ギャンブル	ゲーム	0.201	0.582	1.429
投資	社会	0.263	0.654	1.379
ファッションアクセサリー	投資	0.237	0.648	1.611
コーヒー・ソフトドリンク	電子製品	0.216	0.583	1.415
社会	ギャンブル	0.211	0.446	1.290

表 4.10 信用度 C のみの結果

以上より今回のクラスタリングを行ったうえで、得られたクラスターごとに相関ルール分析を行った試みは、ある程度ではあるが信用度と関連する傾向を捉えられていると言える。しかしながら、信用度 C のみでの相関分析では（オートバイ）を含むルールが抽出されたが、信用度 A のみでのでの相関ルール分析ではそれが見られなかったことなどは、今回の試みからでは捉えられていない傾向が存在することも確かである。

4.3 まとめと考察

本研究では、主成分分析、クラスタリング、相関ルール分析という三つの手法によって分析を行った。クラスタリングからは、極端に SNS をよく利用しているか利用していない人は信用度が低いことや、SNS の利用頻度と年齢とは関係があることなどの大域的な傾向が得られ、相関ルール分析からは、学校・教育関連の投稿と信用度とは関係があることなどのより具体的な傾向を得ることができた。したがって、ある程度ではあるが個人の SNS への投稿に関するデータと信用度とが結びついており、これを信用スコアリングモデルに利用しうることが示唆された。

解釈可能性の観点からは、本研究で用いた分析の流れは正解ラベルである信用度を最後に与える、または与えないことも可能ではあるため、教師データや

新規のデータのバイアスを測定することなどに応用できるかもしれない。また、今回得られた信用度と関わる傾向はせいぜい数個で、さらにその内容は「英語の勉強に分類される投稿をする人は信用度が高い」や「ギャンブルに分類される人は信用度が低い」といったものであり、選択性や事前の信念との一致（確証バイアス）など、人間にとって良い説明の要件をいくつか満たしている。これはローンの審査を受ける側にとっても同じであり、話をごく単純にすると、ギャンブル関連や人間関係の SNS への投稿をやめ、外国語の勉強や健康に関する投稿を意図的にすれば信用スコアは上がり、審査に通りやすくなるかもしれないということである。本研究では、関連研究の議論を踏まえて、説明する対象は資金提供者側を想定しているが、サービスの提供側と需要側の利害が一致しない場合に説明可能性を求めることはゲームの成立を招き、人がモデルを騙す可能性があるため注意するべきであることが確認された。

第5章 総括と今後の課題

5.1 総括

本研究の背景を振り返ると、フィンテックが信用スコアの算出や与信審査のコストを押し下げることにより、開発途上国の金融課題（とりわけ貧困層の金融アクセスの困難性）を解決すると期待されているが、そこには信用スコアリングモデルの解釈可能性という問題が残されていた。そこで、本研究は研究目的を解釈可能性の問題解決に定め、機械学習における解釈可能性の議論を整理し、その社会的な受容性を高めることに焦点を絞って、分析・考察を行い、「個人の SNS への投稿データと信用度がある程度結びついている」という暫定的な結果を得た。しかしながら、「どこまで解釈（説明）できたら信用あるいは納得してもらえるのか」という大きな問題が残されている。「人間にとって良い説明」の項で述べたように、我々の考える良い説明は真実性や確証バイアスの問題により極めて曖昧かつ複雑であるため、説明できたと判断されるポイントが人々の感情や利害関係などによって左右される可能性があるのである。したがって、「どこまで説明出来たら良いのか」という妥協点を見誤ると、与信審査のコストは増大し、前提であった「金融課題を解決する能力」が弱まってしまう。開発途上国においてフィンテック・ビジネスが拡大した背景には、先進国と比べて緩やかな法

規制があったことも忘れてはならない。

以上を考えると、機械学習における解釈可能性の課題は、実世界に適用する際に「精度やコストとのトレードオフ」だけではなく「課題解決能力とのトレードオフ」と捉えるほうが本質的かもしれない。これは開発途上国に限らず、「安心できる AI」を望む先進国においても重要なポイントとなりそうである。これは、トレードオフなのでバランスの問題なのだが、それ自体に正解が存在しないため難解な問題であることに変わりはない。また、郵便番号の識別モデルの例のように、解釈可能性は社会に受け入れられてしまえば必要がなくなるという側面も持っているので、このバランスの問題はさらに複雑になる。

実世界で利用される AI に解釈可能性を持たせるということは、以上の諸点を包括的に議論し、決めていくことなのである。現在、提案されている様々な解釈手法が機械学習を含む AI 技術の研究と社会実装に大きな貢献をすることは間違いないだろう。しかし、これはあくまで手段であり、将来も人が AI の手綱を握っておきたいと考えるならば、様々なケースごとに解釈可能性がどこで、誰に対して必要で、どこまで解釈できたらよいのかという点についての議論を尽くし、その成果を蓄積していくことも重要なのではないだろうか。

5.2 今後の課題

本研究にはいくつかの課題が残されている。まず、クラスタリングについてはサンプルサイズが 562 のデータセットを 8 つのクラスターに分割したために、分析結果の妥当性に関して懸念が残ることとなった。また、相関ルール分析については信用スコアと結びつく特徴的なルールの抽出がうまくできなかったため、データセットの整形や特徴的なルールの抽出に関する他の研究を参考とすることを考える必要がある。

最後に、本研究は信用スコアリングモデルにおける解釈可能性の向上を目的としているが、分析に使用したデータセットの制約から非常に限定された成果しか上げることができなかった。プライバシー保護の観点からは困難であると思われるが、やはりより多くの特徴量を含む大きなサイズのデータセットから信用スコアリングモデルを構築して検証する必要がある。これらの課題は、今後より良いデータセットを入手して解決していきたい。

参考文献

[1] 高野久紀 , 高橋和志, “マイクロファイナンスの現状と課題--貧困層へのインパクトとプログラム・デザイン,” 日本貿易振興機構アジア経済研究所『アジア経済』52 巻、6 号、36-74, 2011.

[2] 岩崎薫里, “東南アジアで台頭するフィンテックと金融課題解決への期待,” 日本総合研究所調査部『環太平洋ビジネス情報 RIM』,Vol.18 No.68, 2018.

[3] 総務省、ＡＩネットワーク社会推進会議, “国際的な議論のためのＡＩ開発ガイドライン案,” 2017.

[4] 内閣府、統合イノベーション戦略推進会議, “人間中心の AI 社会原則,” 2019.

[5] 総務省、ＡＩネットワーク社会推進会議, “ＡＩ利活用ガイドライン～ＡＩ利活用のためのプラクティカルリファレンス～,” 2019.

[6] 原聡, “機械学習における解釈性,” 人工知能学会誌,Vol.33,No.3,pp.366-369, 2018.

[7] 原聡, “説明可能 AI,” 人工知能学会誌,Vol.34,No.4,pp577-582, 2019.

[8] L. Costabello, F. Giannotti, R. Guidotti, P. Hitzler, F. Lecue, P. Minervini , M. K. Sarker, *AAAI 2019 Tutorial On Explainable AI:From Theory to Motivation, Applications and Limitations,* 2019.

[9] C. Molnar, “Interpretable Machine Learning;A Guide for Making Black Box Models Explainable,” 2019.

[10] S. Lessmann, H.-V. Seow, B. Baesensc , L. C. Thomas, “Benchmarking state-of-the-art classification algorithms for credit scoring: A ten-year update, ” *European Journal of Operational Research*,Vol.247,Issue 1,pp124-136, 2015.

[11] 澤木太郎、田中拓哉、笠原亮介、株式会社リコー:研究開発本部、リコーICT 研究所、AI 応用研究センター, “機械学習による中小企業の信用スコアリングモデルの構築,” 人工知能学会研究会資料 SGN-FIN-019, 2017.

[12] 濱田美紀、金京拓司, “財務指標による ASEAN 商業銀行の特徴の分析,” 三重野文晴編『*変容する ASEAN の商業銀行*』, アジア経済研究所、独立行政法人日本貿易振興機構, 2020, pp. 68-75.

[13] 亀岡瑞、船山貴光、宗像昌平、山田実俊、八木圭太、山本義郎, “条件付きアソシエーションルールによる顧客の購入特徴の抽出,” 計算機統計学,Vol.29,No1,pp57-64, 2016.

[14] 岡田孝、元田浩, “相関ルールとその周辺,” オペレーションズ・リサーチ : 経営の科学,Vol.48,No9,pp565-571, 2002.

[15] 伊藤晃, 吉川大弘, 古橋武, 池田龍二 , 加藤孝浩, “アソシエーション分析

における可視化を用いた興味深いルールの探索,”日本知能情報ファジィ学会 ファジィ システム シンポジウム 講演論文集,26(0), pp157-157, 2010.

謝辞

本研究を行うにあたり、主指導教員である北陸先端科学技術大学院大学知識科学研研究科 Dam Hieu Chi 教授には、ゼミを通して丁寧なご指導を頂くなど、大変お世話になりました。この場を借りて感謝を申し上げます。

また、本研究を進める過程で多くの示唆をくださった Dam 研究室のメンバー、また合同で勉強会を行った際に、多くの助言や示唆をくださった郷右近秀臣先生および郷右近研究室の皆様にもこの場を借りて感謝を申し上げます。