人工知能学会第二種研究会資料
市民共創知研究会
SIG-CCI-012-01

山陽小野田市の若者の転入出要因分析

Young Migration Factor Analysis of Sanyo-Onoda City

高田 寛之[1*]　山本 吉紀[2]　藤田 晋礼[2]　井上 啓[2]
Yoshiki Yamamoto[2]　Akinori Fujita[2]　Kei Inoue[2]

[1] 山陽小野田市立山口東京理科大学 工学部 数理情報科学科
[1] Sanyo-Onoda City University, Faculty of Engineering,
Depeartment of Informatics and Data Science
[2] 山陽小野田市立山口東京理科大学 工学部 電気工学科
[2] Sanyo-Onoda City University, Faculty of Engineering,
Department of Electrical Engineering

Abstract: We specify (young) migration factors from the anonymous survey of Sanyo-Onoda City in 5 years (2018-2023). The analysis is based on machine learning and shapley value analysis. The inflow factor candidates are “Relocation: Relative are nearby”, “We have a owned house.”, “Relocation: Convenient transportation for commuting to work and school, etc.”, “Relocation: We have land or a house.”, “Relocation: There is a place of employment (workplace) .”, and “Relocation: Good living/housing environment”. Whereas, as the outflow factor candidate is “Relocation” with other reasons.

1　はじめに

本研究の目的は、山陽小野田市の若年層世代の転入出の要因特定である。本来の目的は、山陽小野田市の出生率（人口千人あたりの出生数）が、特に近隣自治体よりも著しく下がっている原因をデータから調査してほしいという要望に応えることである。ところが、手元にある分析に耐えるデータは転入出時の匿名アンケートであり、出産に関する意識調査アンケートではない。そこで、研究目的を出産世代である若年層の人口流出を食い止めるための政策に資する情報の提供に読み替えた。

本研究の背景・動機となる山陽小野田市の出生率減少問題について説明する。近年の山陽小野田市の出生率は、近隣自治体と比較して減少が著しい。出生率減少は、現行の年金制度の前提を覆し、制度の再設計が必要になるだけでなく、社会基盤の再設計も必要になる。これらの再設計にはかなりのコストがかかるため、山陽小野田市としては現行体制を維持できるように限られた財源の中で、効率良く、ピンポイントに財源を投入する必要がある。ところが、これまでの政策の効果が出ているとは言い難い現状とのことであった。そこで、我々が相談を受け、分析をすることになった。

*連絡先：山陽小野田市立山口東京理科大学
工学部 数理情報科学科
〒 756-0884 山口県山陽小野田市大学通 1-1-1
E-mail: htakada@rs.socu.ac.jp

予備実験として、市から提供を受けた異動データと匿名アンケートの接続を試みた [1]。理由は、目的変数と定めた期待出生率は異動データの情報を使って計算できるものであり、一方で説明変数の大部分は匿名アンケートにあったからである。接続がうまくいかない理由は、異動データと匿名アンケートをリレーショナルデータベースと見たときに、主キーがかなりの数のレコードにおいて欠測しているからである。そもそも匿名アンケートは出生率減少の分析を目的としたものではなく、人口の転入出の分析のために作られていた。

一方で、予備実験として、異動データを元に山陽小野田市から近隣自治体（宇部市、下関市、山口市など）への人口流出についての仮説検定を行った [2]。確かに宇部市や山口市への人口流出は有意差が見られたが、下関市に関しては有意差があるとは言えなかった。また山本 [2] は、クラスタリングアルゴリズム（K-means++）を利用して、山陽小野田市と同規模の世帯数を持つ自治体の中から、出生率の増加傾向の意味でのランキングを作成した。（熊本県合志市、愛知県常滑市、石川県野々市市、岐阜県瑞穂市、愛知県みよし市は上位）このランキングは、規模の大きい自治体は元々人口が集まりやすい何かがあってそれに抗う政策をうつことは難しく、同程度規模の自治体でないと真似できる政策は少ないのではないかという仮説によるものである。なお、分析に用いたデータは、e-Stat[3] から取得できる。

本研究の方法論は、前処理（これらの機械学習ツールに合う形にデータを整形すること）、機械学習ツールを適用して結果を得ることである。更に研究目的を達成するにはドメイン知識や因果分析と合わせて政策で操作可能な原因を特定する必要があるが、これは市の担当者などと協力しながら実行する必要があるので、本研究では、その手前の段階である、相関の高い変数の列挙までを行う。

本研究の貢献は、山陽小野田市の全年齢層及び若年層の転入出の要因の候補が得られたことである。

本論文の残りは、4 章からなる。2 章では、データに関する説明、分析方法、結果の読み方について説明する。3 章では、実験結果をまとめる。4 章ではまとめと今後の課題について述べる。

2 方法

アンケートのように表として与えられたデータの要因分析でよく用いられている方法やそのツールについて紹介する。アンケートデータに目的変数を予想するための情報が説明変数にそもそもあるかを調べるには、PyCaret[4] のような AutoML ツールを用いると良い。これは、少ないコードで、様々な機械学習モデルの比較、学習とチューニング、識別・予測性能や特徴量の重要性のグラフ表示ができる。特に情報の有無は AUC を観察するとわかる。目的変数の各値に影響を与える説明変数を調査する方法は、例えば協力型ゲーム理論の結果に基づいた Shapley 値を観察する方法がある [5]。文献 [6] は、Shapley 値についての解説もわかりやすいが、それ以外の機械学習モデルに関する説明技術についても詳しい。Shapley 値のライブラリのドキュメントに指摘されている通り、この方法は本質的に相関分析と同等であるので、即座に政策へ反映させようとすることは控えるべきである。列挙された変数は目的変数と偽相関があるだけなのかもしれないので、原因として特定するには、更にドメイン知識や因果推論・因果探索 [7] を使って因果分析を行ってから、政策に反映させるべきである。

なお、分析の大まかな手順は、以下のように行う。

1. 前処理
2. アンケートデータにそれなりの情報があるとみなせるか？の確認。AUC が 0.7 以上達成できる程度に説明変数が目的変数を識別・予測できる学習モデルが存在するか？その学習モデルにおいて、どの説明変数が目的変数を説明するのに重要か？
3. 各説明変数が目的変数の値の 0,1 をとるのにどういう貢献をしているか？の確認。xgboost モデルで学習して、Shapley 値の表示を行い、それを読み取る。

2.1 匿名アンケート

ここでは、分析する対象であるデータの元になった匿名アンケートの取り方やアンケートの設問について述べる。

アンケートは転入出などで役場に手続きに来た人を対象に、匿名で記入してもらったものである。データは 2018 年から 2022 年の 5 年分あった。集計状況はそれぞれの年で転入と転出と転居（市内間異動）に分けて集計されていた。また異動データと比べると、アンケート記入者がたとえ若年層でなくても、世帯主が 50 代（父母）でその子供夫婦とその子供（世帯主からの続き柄は子の子）のように出生が期待できる人が大きな世帯に埋もれていることもあるため、単純に世帯主の年代だけに限ってしまうと、大世帯に属する若年層の意見が反映されないこともある。しかしながら、このことを考慮してアンケートが取られているわけではなく、その世帯に若年層が存在するかどうか判断するための情報は無いので、本研究では参考のために全年齢層のデータと世帯主が若年層（２０代と３０代の世帯主が世帯）のレコードのみからなるデータの二種類で分析を行う。

次にアンケートの設問と選択肢の一覧を示す。転入時アンケートと転出時アンケートは、共通した項目の問が 6 つある。一覧は実物のアンケートと問の番号や選択肢の前の記号に若干の違いがあるが、便宜的なものである。

問 1 異動した主な理由（いずれか一つ選択）

R1 転勤、**R2** 就学、
R3 出産や療養などに伴う一時的な転入、
R4 就職・転職、**R5** 創業、**R6** 婚姻など、
R7 住み替え、**R8** その他（自由欄）

問 2 問 1 において 7 住み替えを選択した人向けに住み替えを決めた理由（当てはまるもの 3 つまで選択）

R701 転入奨励金制度がある
R702 雇用の場（職場）がある
R703 商売や事業経営がしやすい
R704 通勤・通学などの交通の便が良い
R705 親族が近くにいる
R706 生活・住宅環境が良い
R707 医療・福祉面が充実している

R708 子育て・教育環境が良い

R709 買い物・娯楽などの場が多い

R710 余暇や生きがいを楽しむ場が多い

R711 土地・家がある

R712 その他（自由欄）

問 3 異動人数（男と女の内訳も含む）

問 4 異動した人の代表者の年代 １０代、２０代、…、６０代、７０代以上

問 5 住まいの形態（いずれか一つ選択）

H1 持ち家（一戸建て）

H2 持ち家（分譲マンション）

H3 借家（一戸建て）

H4 民間賃貸住宅（アパート、賃貸マンション）

H5 会社の寮や社宅

H6 公的賃貸住宅（県営住宅、市営住宅等）

H7 その他（自由欄）

問 6 山陽小野田市への転入者数を増やすために、市が実施した方が良い効果的な支援はどれだと思うか（あてはまるもの 3 つまで選択）

M01 家を借りる際の家賃補助（期間限定）

M02 空き家を購入する際の補助

M03 住宅をリフォームする際の補助

M04 固定資産税の減免（期間限定）

M05 高齢者や障がい者の方が入居しやすい住宅の普及

M06 親族の近くに住むことや同居に対する支援

M07 子育て世代向け公的賃貸住宅の供給

M08 子どもの医療費補助　**M09** 出産祝い金

M10 保育費支援　**M11** 学校授業料支援

M12 転入奨励金　**M13** 新規就農者への助成金

M14 起業支援・助成　**M15** その他（自由欄）

2.2 前処理

目的変数は、転入 1 転出 0 とした。問 2, 問 3, 問 6, 問 8 は選択肢の選択の有無に応じて 1,0 を割り振り、カテゴリカル説明変数とした。問 4 の異動人数と男女の内訳は整数値として、数量説明変数とした。なお 0 人が NaN になっていたのでその修正を行った。問 5 の年代情報は「代」や「代以上」の文字を消去して整数値に直し、数量説明変数とした。

全年齢層データについては、全てのレコードを残し、若年層データについては年齢が 20,30 のレコードだけを残し、他の年齢のレコードは消去したものを利用した。全年齢層のレコード数は 3253 個、若年層データのレコード数は 1970 個であった。train と test の割合は 8:2 とした。多重共線性の基準は 0.7 に設定し、多重共線性があるとみなされる説明変数は目的変数との相関が高い方を残し、それ以外は削除するように設定した。

2.3 転入出を説明する情報の有無

アンケートに転入出を説明する情報があるかの分析手順を述べる。目的変数 io が転入 1 と転出 0 の 2 値をとるカテゴリカル変数であるので、PyCaret の識別タスクライブラリを用いた。AUC の意味で最も良いモデルの ROC 曲線または AUC（ROC 曲線が作る図形の面積）を見ることでアンケートに十分な情報があるかわかる。情報がなければ ROC 曲線は $y = x$ の直線またはそれを下回り、情報があれば、$y = \sqrt{x}$ のような形状の曲線となる。AUC は理屈の上では最大で 1 であるが 0.999 などの場合は、何かしら情報リークを疑う必要がある。AUC が 0.7 から 0.85 ぐらいだとアンケートに情報があるとみなせる。またハイパーパラメータチューニングして、識別精度が十分ある識別器を作成すると、重要な特徴量の一覧を重要度と共に得ることができる。

PyCaret の手順は以下のとおりである。

1. 指標 AUC に関する複数モデルの性能比較
2. 最も AUC が大きい性能を示したモデルのハイパーパラメータチューニング
3. AUC の確認（ROC 曲線の表示）
4. 重要な説明変数の一覧取得

次に示すのは、比較する学習モデルと、その省略記号である。

gbc Gradient Boosting Classifier

ada Ada Boost Classifier

lr Logistic Regression

lda Linear Discriminant Analysis

xgb Extreme Gradient Boosting

rf Random Forest Classifier

qda Quadratic Discriminant Analysis

et Extra Trees Classifier

nb Naive Bayes

dt Decision Tree Classifier

knn K Neighbors Classifier

dmy Dummy Classifier

svm SVM-Linear Kernel

ridge Ridge Classifier

2.4 目的変数の各値に貢献する説明変数

目的変数の各値に貢献する説明変数を見つけ出す方法として、本研究では Shapley 値を用いる。他にも [6] では、機械学習モデルを説明するための技術がまとめられていて Shapley 値以外にも目的変数に貢献する説明変数を調べる方法がある。Shapley 値を採用した理由は、単にライブラリの使いやすさと貢献値の考え方が単純で理解しやすかっただけであり、科学的な良し悪しを吟味した上での選択ではないことを断っておく。

Shapley 値の概念について説明する。Shapley 値は、協力型ゲームにおける概念である、平均限界貢献度として知られている。これを理解するためには、平均を取る前の限界貢献度について理解する必要がある。A さんと B さんが共同作業したときに得られる報酬が 20 万円であり、個別に働いた場合は、A さんは 10 万円、B さんは 6 万円の報酬を得るような状況を考える。A さんにとっての限界貢献度は、A さんが参加したか否かにおける報酬の差額として定義される。ただ、B さんが加わっているか否かで状況が異なり、限界貢献度は状況毎に値が異なることがある。可能な状況に関して限界貢献度を求め、その平均値を Shapley 値と呼ぶ。例えば、A さんの限界貢献度を 2 つの状況（B さんの参加状況）によってそれぞれ求める。B さんがいなかったとき、すなわち A さんが単独で働いたときと A さんが働かないときの差額は 10 万円なので限界貢献度は 10 万円である。B さんだけが働いている状況と二人で共同作業する状況での差額は、20 万円 - 6 万円＝ 14 万円であるから、この場合の A さんの限界貢献度は 14 万円である。従って Shapley 値は $(10+14)/2 = 12$ なので 12 万円である。同様に B さんの限界貢献度は、A さん不参加の状況では、6 万円 - 0 万円=6 万円、A さん参加の状況では 20 万円 - 10 万円 = 10 万円なので B さんの Shapley 値は $(6+10)/2 = 8$ より 8 万円である。Shapley 値は一緒に作業したときの報酬の取り分として、個人の能力が考慮された公平な分け方であることが知られている。

Shapley 値の意味は一緒に働いたときの報酬 20 万円を A さん 12 万円、B さん 8 万円と配分するのが妥当という意味である。もちろん 10 万円ずつ配分するという考え方もあるが、この配分は単独で働いたときの重みが加味されていない。

機械学習の説明への適用は次のような対応付けを行う。先の例における A さん、B さんは説明変数 A、B に読み替え、報酬はその説明変数をいれたときといれないときの識別性能の差に読み替える。識別性能の貢献度として Shapley 値が使える。

Shapley 値のグラフは、その変数が大きい値をとったときと小さい値をとったときで色が異なる。その変数が目的変数 1 に寄与する場合は、1 の方向に赤い点がプロットされる。逆に寄与しない場合、目的変数 0 に寄与する方向に値がずれる。

本研究の実験では xgboost モデルで学習して、python の shap ライブラリを使用して、結果を出力した。

3 結果

ここでは、全年齢層のデータと若年層のデータについての分析結果を述べ、その解釈を説明する。

3.1 PyCaret による結果

PyCaret で全年齢層データと若年層データのそれぞれのデータにおいて様々な学習モデルを適用し、AUC の高い順に並べた表をそれぞれ表 1 と表 2 に示す。このデータに対して最も AUC が高かったのは gbc: Gradient Boosting Classifier （勾配ブースティング分類モデル）であった。

表 1: AUC を基準にしたモデル比較: 全年齢層: TT の単位は [ms]

model	Acc.	AUC	Rcall	Prec.	F1	TT
gbc	0.73	0.82	0.73	0.74	0.72	74
ada	0.71	0.80	0.71	0.72	0.71	68
lr	0.72	0.80	0.72	0.72	0.72	198
rf	0.72	0.79	0.72	0.72	0.72	103
lda	0.72	0.79	0.72	0.72	0.71	53
et	0.72	0.78	0.72	0.72	0.72	94
qda	0.64	0.77	0.64	0.79	0.58	550
nb	0.64	0.74	0.64	0.79	0.58	151
dt	0.68	0.69	0.68	0.68	0.68	146
knn	0.63	0.68	0.63	0.64	0.63	158
dmy	0.52	0.50	0.52	0.27	0.36	52
svm	0.66	0.00	0.66	0.72	0.62	56
ridge	0.72	0.00	0.72	0.72	0.71	52

そこで gbc に対して、ハイパーパラメータチューニングを行い（イテレーション回数は 150, 交差検証は 10 分割）学習済みモデルの性能を調べた。図 1 と図 2 の

表 2: AUC を基準にしたモデル比較: 若年層 TT の単位は [ms]

model	Acc.	AUC	Rcall	Prec.	F1	TT
gbc	0.72	0.79	0.72	0.74	0.70	63
ada	0.72	0.77	0.72	0.73	0.70	62
lr	0.72	0.77	0.72	0.73	0.70	188
lda	0.72	0.77	0.72	0.73	0.70	50
xgb	0.70	0.76	0.70	0.70	0.70	59
rf	0.71	0.75	0.71	0.70	0.70	99
qda	0.71	0.74	0.71	0.80	0.65	50
et	0.71	0.73	0.71	0.71	0.70	91
nb	0.70	0.70	0.71	0.80	0.65	145
dt	0.68	0.67	0.68	0.68	0.68	144
knn	0.64	0.65	0.64	0.63	0.63	152
dmy	0.59	0.50	0.59	0.35	0.44	48
svm	0.64	0.00	0.64	0.75	0.58	52
ridge	0.72	0.00	0.72	0.74	0.70	49

ROC 曲線を見てわかるように $y = x$ の線より上方向に ROC 曲線が描かれている。これはアンケートが転入転出を説明する情報を持っていることを表す。実際に、図 3, 図 4 の混同行列を見ても 0,1 の予想が正解クラスの 0,1 を言い当てている傾向が見える。

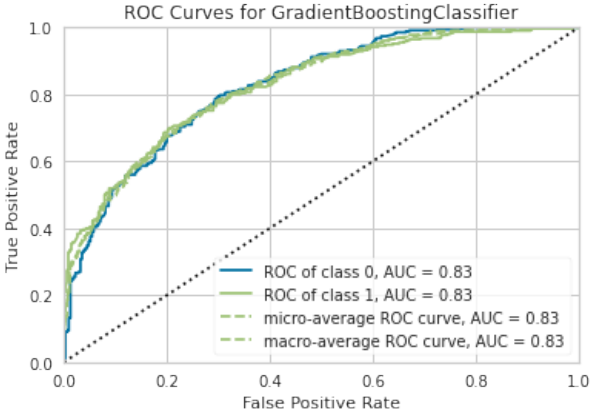

図 1: 勾配ブースティング分類の ROC 曲線:全年齢層

それなりの識別精度を持ったモデルにおいて、どの説明変数が貢献しているかを特徴量の重要さのグラフ図 5 と図 6 から調べよう。図に表示されている特徴量は上から順に重要度が高く、横軸はその変数の重要度を表している。各特徴量はアンケートの選択肢である。英語表記にしているのは文字化けを防ぐために英語表記に直したためである。読者が読み取る際は R705 などのコードを 2 節に述べた選択肢と照らし合わせると日本語の選択肢がわかる。age, total, male, female はそれぞれ年代、異動時人数、異動時人数の男性の人数、

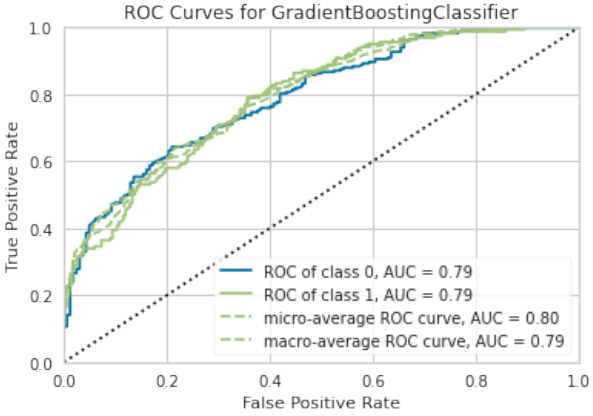

図 2: 勾配ブースティング分類の ROC 曲線:若年層

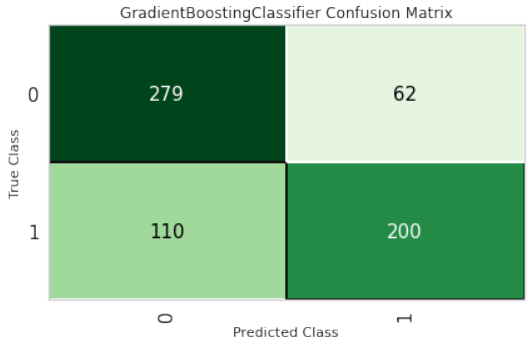

図 3: 混同行列:全年齢層

異動時人数の女性の人数である。

図 5 には R2 School が上位にあるが、図 6 には見当たらないのは、後者が 20 代 30 代のみに制限したため、就学による 10 代の転入のデータが除去されたためである。重要度の中で 0.08 以上の指標は R705（親戚が近くにいる）, が両方の図にあり、若年層に限ると R704（交通の便が良い）となっている。0.04 以上にするといくらか指標が増え、大体上位の特徴量の顔ぶれは似ている。

3.2 Shapley 値の結果

特徴量の重要度は転入に効いているのか転出に効いているのかはこのままではわからないので、Shapley 値の結果を見て判断する。xgboost の学習済みモデルの shapley 値をプロットしたものを図 7 と図 8 に示す。図の読み方は、上から順に重要な特徴量が並んでいて、その説明変数が赤いほど大きな値をとり青いほど小さな値をとっている。0,1 のカテゴリカル変数の場合は 1 のとき赤で 0 のとき青である。また横軸の正の方向に伸びているとその変数が転入に貢献していることを表し、

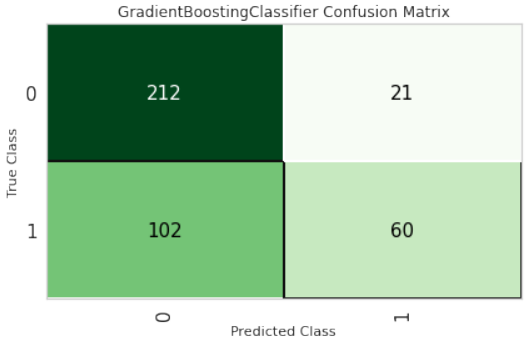

図 4: 混同行列:若年層

負の方向に伸びているとその変数が転出に貢献しているとみなせる。

重要な変数は、モデルが変わっても、多少順位変動はあるが、顔ぶれはあまり変わらないことがわかる。

図 7 について、赤い点が右にたくさん伸びている変数は、R2 就学、R705 親族が近くにいる、H1 持ち家（一軒家）、R711 土地・家がある、R704 通勤・通学等の交通の便が良いなどが目立つ。逆に左方向に向いている変数は転出に効いている変数である。R7 住み替え、M7 子育て世代向け公的賃貸住宅の供給、M9 出産祝い金、M3 住宅をリフォームする際の補助が、若干左に伸びているように見える。

図 8 について、赤い点が右に伸びている変数は R705 親族が近くにいる, H1 持ち家（一軒家）、R704 通勤・通学等の交通の便が良い、R711 土地・家がある、R702 雇用の場（職場）がある、R706 生活・住宅環境が良いが上げられる。逆に左方向に伸びたものは R7 住み替えにつけているもので、右に伸びている要因以外の理由によって転出している傾向がうかがえる。

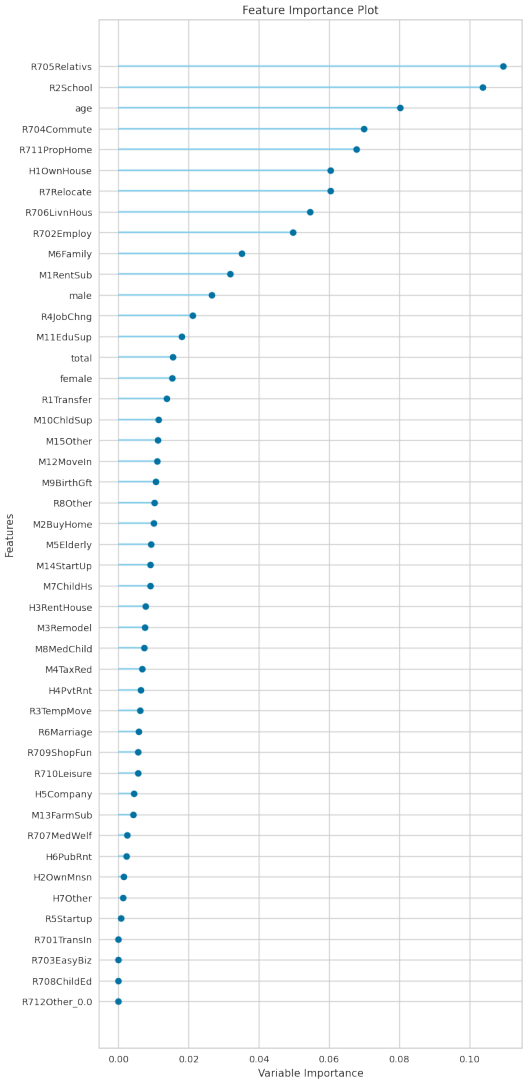

図 5: 重要な特徴量:全年齢層

4　おわりに

本研究では、転入出匿名アンケートと機械学習モデルと Shapley 値を用いて、どの選択肢を選んだ人が転入、転出のアクションを起こしているかの相関関係を調査した。結果としてわかったことは、若年層が世帯主の世帯は、親戚が近くにいる、持ち家（一軒家）がある、通勤通学の交通の便が良い、土地・家がある、雇用の場（職場）がある、生活・住宅環境が良いという理由で転入しているのに対して、転出理由はそれ以外の様々の要因で住み替えという形でアクションが起きていることがわかった。また全世帯層で観察すると、これ以外の転出理由として、子育て世代向け公的賃貸住宅の供給、子どもの医療費補助、住宅をリフォームする際の補助が不満である傾向が伺えた。

今後の課題として、ドメイン知識や因果推論・因果探索の手法を用いて、相関分析ではなく因果関係を特定して、政策立案のために要因の信頼性を上げたり、出生率上昇のためのアンケートを取り直すなどして、出生率向上のための政策立案のための信頼性の高い情報の提供に努めたい。

謝辞

本研究の実施にあたり、データおよびドメイン知識の提供いただいた山陽小野田市企画課の松岡祥吾様、藤井貴大様に深い感謝の意を示す。また、データ提供に関わった行政の皆様、市民および元市民の皆様に深く感謝の意を示したい。

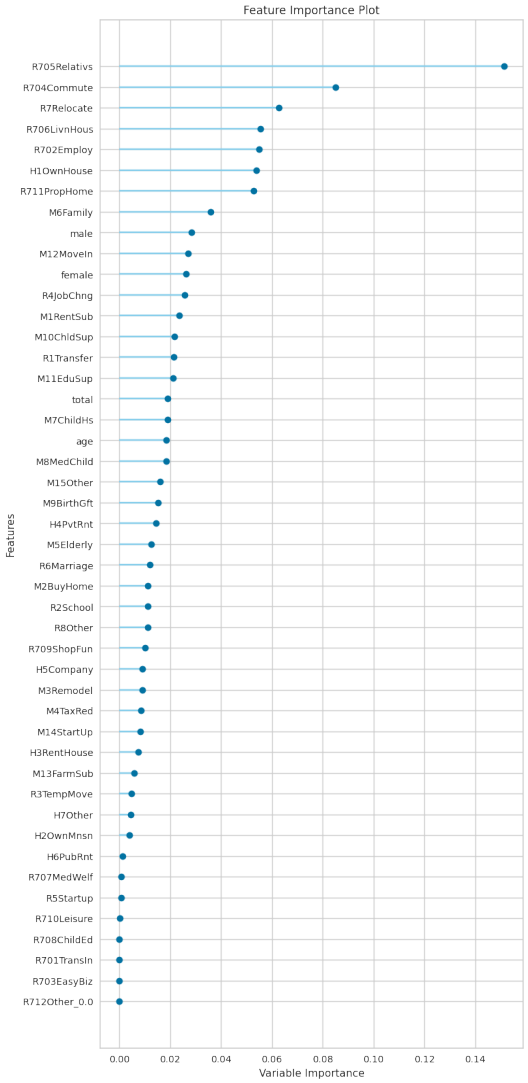

図 6: 重要な特徴量:若年層

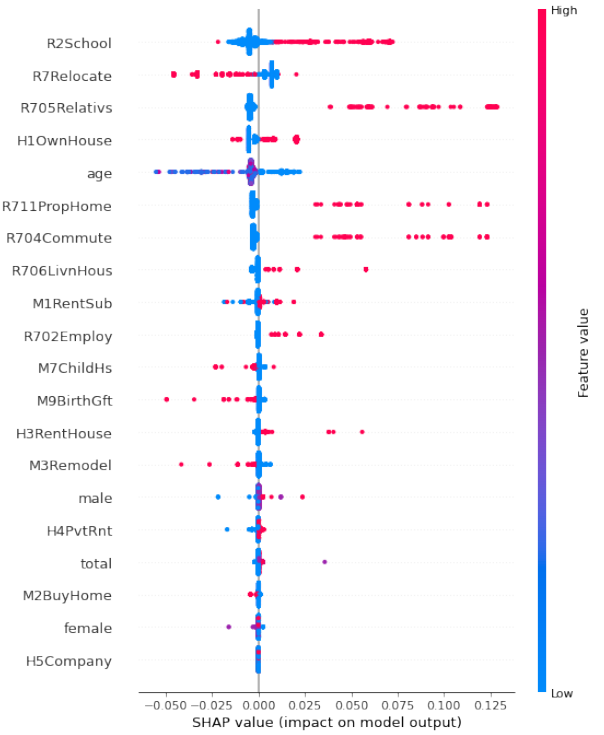

図 7: 全年齢層の転入出に関する shapley 値

参考文献

[1] 藤田晋礼: “相関分析による出生率低下因子の探求：山陽小野田市データを用いた事例研究”, 卒業論文, 山陽小野田市立山口東京理科大学, 46pages (2023)

[2] 山本吉紀: “仮説検定とクラスタリングを用いた山陽小野田市の出生数減少の分析”, 卒業論文, 山陽小野田市立山口東京理科大学, 33pages (2023)

[3] e-Stat 政府統計の総合窓口 （トップページ/地域から探す/社会・人口統計体系/ 地域ランキング（市区町村データ）） https://e-stat.go.jp

[4] Moez Ali, “PyCaret: An open source, low-code machine learning library in Python, https://www.pycaret.org (2020)

[5] Lundberg, Scott M and Lee, Su-In, “A Unified Approach to Interpreting Model Predictions”, *Advances in Neural Information Processing Systems 30*, Curran Associates, Inc. pp.4765-4774 http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf(2017)

[6] Christoph Molnar: *Interpretable Machine Learning* https://christophm.github.io/interpretable-ml-book/ (2019)

[7] 高橋将宜, 統計的因果推論の理論と実装, 共立出版 (2022)

図 8: 若年層の転入出に関する shapley 値