機械学習を用いた選挙報道における有用な情報提示方法の考察

○大倉清司　浅井達哉　岩下洋哲　垣渕太成　福田茂紀（富士通）
大堀耕太郎（東洋大学）

A Study on Providing Valuable Information in Election Coverage by Machine Learning

* S. Okura, T. Asai, H. Iwashita, T. Kakibuchi, S. Fukuta (Fujitsu Ltd.)
and K. Ohori (Toyo University)

概要－　近年，国政選挙の投票率が低くなっている．選挙当日に報道される開票特別番組において，当確を早く出すことばかりでなく，政治に興味を引くような，当落理由の提示のような有権者にとって有用な情報が望まれる．本研究では，報道という観点からこの問題へ対処するため，有権者にとって有用な理由を説明変数の組み合わせで重要度の高い順に出力する機械学習の活用を検討した．この際，出力する理由の数や理由を構成する説明変数の数とその有用性の関係が不明である．そこで本研究では，提案する機械学習手法において，この関係について定量的に検証し，重要な理由ほど有用な情報を提示できている傾向を明らかにした．また，政治学で研究されているが新聞記事に含まれないような新規の理由を提示できている割合を調査し，機械学習で理由を出力する有効性を示した．今後は有権者の視点での有用性を，被験者実験を通じて明らかにしていく．

キーワード:　機械学習，選挙，予測，理由，有用性

1　はじめに

国政選挙の開票日に，各テレビ局が一斉に開票速報の特別番組を放送するのが通例となっている．テレビ局をはじめメディアの目的は，高い視聴率を取ることであるが，選挙の投票率や政治に関わるテレビ報道が減少している [1)2)3)] ことから，政治に興味がある有権者が減っていると考えられる．有権者に興味深く有用な情報を提供し，報道の質を上げることが求められている．

しかし報道には制約がある．制作にはコストの制約があるため [4)]，十分な取材を行い，有権者の興味を引く当落要因について解説できない．また，放送上のルールとして，開票日以前に有権者にとって有用な当落予測を提示することは，公職選挙法により禁止されている(公職選挙法第 138 条の 3　(人気投票の公表の禁止))．また，放送時には倫理に基づいた情報提示が必須である [5)6)]．

開票速報番組においては，上述した制約のため，当落理由の説明など，有権者が選挙を深く理解するための有用な情報よりも，投票締め切り後にいかに早く当選確実を伝えるかという出口調査に頼った早打ち競争中心の報道になってしまっている．若者は，このような報道姿勢を冷めた目でみていることが指摘されている [7)]．選挙後の新聞報道においても，紙面の制約から選挙区に着目した深い分析をすることが難しく，特に注目された選挙区中心の解説にとどまっている．

本研究では，開票後の選挙報道において，　機械学習を活用することにより，コストをかけず，放送倫理などのルールを守った上で，有権者にとって有用な当落理由に関する情報を報道する方法に焦点をあてる．具体的には，メディアの持つ候補者のプロフィールや選挙区情報，政党情報などから，機械学習技術を用いて候補者の当落についての予測とその予測理由を提示することで，人手による労力をかけることなく，自動で有権者をひきつける報道を実現することを目指す．

本研究と同様に，機械学習を用いて選挙予測を行い，その理由を分析する研究がいくつかある [8)~13)]．これらの研究においては，機械学習により報道にも耐えうる比較的高い精度での当落予測を実現できているが，その予測理由に関する情報が選挙報道に有用であるのかを考察することはできていない．ここで選挙報道における情報の有用性とは，有権者の投票行動に役立つことと定義する．

報道においては，決められた放送時間や記事の分量の下で，この有用な情報を選択する必要がある．機械学習によって出力される当落要因の説明の中には，過去の選挙で多く当選している政党の候補者が有利，というような，有権者にとって既知のデータ傾向を示しているような説明も含まれており，これをそのまま提示しても有用であるとは限らない．著者らが機械学習の選挙への適用の取り組みを共同で実施した TBS テレビ [14)] の選挙の専門家からは，例えば世襲候補は強いという傾向はよく知られているが，世襲かつ○○の候補は弱い，といったような，変数属性の組み合わせ如何で判定が覆る意外性のあるケースが有用な情報の例として挙げられた．以上より，単純に多くの説明変数を加えたり，モデルを複雑にすることで予測精度を向上しても，予測に寄与した説明変数が有用な情報につながるとは限らない．これは，選挙に限らず，機械学習の社会実装が進まない要因としても重要な視点である．

そこで，本研究では，当落予測モデルに基づいて提示する情報とその予測理由の有用性の関係を明ら

かにすることを目的とする．ここで，有用性の判断は，大きく 2 つのステップで決まる．まずテレビ局などのメディアにおける選挙の専門家による有用性の判断がある．選挙の専門家の知識に基づいて報道に用いる情報を選択する．次に，報道を通して有権者が情報の有用性を判断する．自身の投票行動に関連付けて，新たな気づきなどを得られると有用だと判断する．本論文では，この第 1 ステップに焦点を当て，選挙の専門家が実際に報道した内容を一定レベルでの有用な情報であると仮定し，機械学習から提示された情報を評価する．実際，選挙の専門家による解説記事は，有権者にとって有用であるという視点でかかれており [15]，新聞記事中に記載される情報は，有用である可能性が高い．そこで，本研究では，新聞記事中に出現するキーワードにより有用性を定量評価する．

2 関連研究と本論文の研究課題

選挙において，機械学習を使った予測についての研究は多数あるが，精度中心の議論となっている．予測精度だけでなく，予測理由の説明に焦点をあてている従来研究は 2 つある．

Hummel ら [8]は，様々な「基礎データ(fundamental data)」（政治経済指標，候補者のプロフィール，支持率など）を使い，アメリカ大統領選挙，上院議員選挙，知事選挙の当落予測モデルをプロビット回帰分析を用いて構築した．この手法の大きな利点として，費用と時間がかかる得票直接調査が不要であること，その結果，候補者さえわかれば，選挙に先立って予測が可能なことが挙げられる．彼らは高精度なモデル構築に成功している．この中で，経済指標や候補者プロフィールがどのくらい意味がある変数かということを分析している．また，経済変数や大統領支持率などの３Ｑのデータは，１Ｑ,２Ｑのデータに加えても大して変わらないことを分析している．それぞれの選挙の違いも分析している．彼らは，過去の選挙の経験的な証拠をもとに，複数のダミー属性を追加することで，1980 年～2012 年の全選挙のうちの 90.2%（within-sample），89.1%（out-of-sample）を当てている．また，回帰分析により，どの説明変数が重要かどうか，変数ごとに重要度を出すことができ，人間が解釈しやすい形で出力できている．この研究のように，線形回帰のモデルでは，予測精度の向上のために，変数を追加していく中で重要な変数が変化していく．

Rusmawati[9]は，オントロジーベースの決定木と，自動推論機能を備えた機械学習によって，2019 年インドネシア議会選挙の予測理由の解釈を試みている．10-fold の交差検証の結果，予測精度はそれぞれＦスコアで 0.83, 0.86 となっており，高い精度を実現している．ここで理由とは，モデルから出力された，属性と閾値の条件の組み合わせである（例：以下）．

F1: (Party ∈ Demokrat, Gerindra, Golkar, NasDem,
<?TeX PAN, PDIP, PKB, PKS, PPP} ?>
<?TeX ∧ (APM ≥ 3.2190e + 01) ∧ (IPM ≤ 8.3060e + 01) ?>
<?TeX ∧ (Latitude ≥ -9.8500e + 00) ∧ (Latitude ≤
4.2399e + 00) ?>
∧ (Longitude ≥ 9.6530e + 01) ∧ (Longitude ≤ 1.3815e + 02)
∧ (PPM ≥ 3.4200e + 00) ∧ Result ∈ {0, 1} ∧ Gender ∈ {F}
∧ (IPM ≥ 6.3758e + 01) ∧ Resident ∈ {Yes} ∧ (CandiNum ≥ 2.0019e + 00)
∧ (PPM < 6.6397e + 00) ∧ (APM < 5.6768e + 01)
∧ (CandiNum < 3.0039e + 00))

彼らは，ツールを用いて，意思決定を定式化でき，インスタンスレベルでチェックできるとしている．しかし，スケーラビリティの問題にも触れており，条件の組み合わせ式が長くなるため，注意深い解釈が必要としている．一般的に，得られた組み合わせ式から機械学習の予測モデルの精度結果に結びつく条件の種類を理解できるとしている．理由は複数の説明変数の組み合わせで出力されており，視聴者が知らない知識を出力していると言える．しかし，候補者の当落の予測理由としては，１つの決定木しかなく，重要度という点では示しにくい．説明変数の組み合わせは有用であるものの，報道に活用するには壁がある．この研究のように，決定木による予測モデルでは，予測精度を向上させるために変数を加えていくと多数の変数を組み合わせた木構造のルールができあがる．

以上のように，予測精度向上を目指して説明変数を増加させると，説明に用いられる説明変数，あるいは説明変数の組み合わせが変化する．そこで，本論文では，Hummel らの手法と Rusmawati の手法の両要素を扱い，説明変数の組み合わせを線形回帰により予測するモデルを用いて選挙報道に有用な情報を提示する状況を想定する．具体的には，説明変数の組み合わせを 1 つの理由として，各理由に重要度を付与し，その重要度の順に選挙報道で用いるべき情報を自動的に提示する．

そして，本論文の研究課題は，本モデルによって理由の数や理由を構成する説明変数の組み合わせの変化により，報道においてどの程度有用な情報を提示できるのかを示すことにある．

既存研究の手法と本研究の手法の位置づけを Table １にまとめる．本実験の目的は，我々の手法（従来よりも詳細な理由を複数出力）で出力した理由のリストは，従来の報道で用いられている有用な理由がどれくらい含まれているかを定量的に示すことである．有用性について定量的に示すための指標を提案し，それに基づいて考察を行う．

Table 1：既存研究と本研究の位置づけ

手法	機械学習モデル	理由の種類	出力する理由の数
Hummelら[8]	線形回帰	単一の説明変数	複数
Rusmawati[9]	決定木	説明変数の組合せ	１つ
本研究	ルール線形モデル	説明変数の組合せ	複数

3　実験方法

本章では，研究課題に答えるために用いる実験データ，当落予測モデル，実験対象とする選挙区，実験について説明する．

3.1　実験データ

今回対象とするのは，日本の衆議院選挙小選挙区の候補者当落の予測問題である．過去 3 回分の衆議院選挙（2012，2014，2017 年の選挙）のデータを使って学習し，2021 年衆議院選挙の予測モデルを構築する．候補者に焦点を当てた予測理由の説明をするため，従来研究にならい，選挙の候補者を行，各候補者の属性を列とする表形式データを作成し，各候補者の当落をラベルとしてデータ化する．候補者の属性については，過去の政治学における分析 [16)～28)] をふまえ，当落に影響があるとされるものや，政党支持率などの情報をもとに決定した（Table 2）．なお，都議会選挙の衆議院選挙への影響を示唆する研究 [29)]や，衆参議院の補欠・再選挙が政党にとって大きな影響があることを示唆する解説がある [30)]ことから，都議会選挙や衆参議院補欠・再選挙についての情報も含めた．候補者や政党の政策については，複数回の選挙を通じての共通の属性としてデータ化する必要があるが，例えば消費税増税，新型コロナウィルスへの対策など，選挙ごとに異なるものも多く，今回は対象外とした．

各候補者の属性については総務省のサイトと Wikipedia から収集した．政党支持率については，インターネットから収集し，集計した．地盤の属性については，明確な定義がないため，インターネッ

Table 2: 使用属性

属性大項目	属性中項目	属性	数値/カテゴリ （）内は数値の離散化閾値	例／2値化後の値の意味	新聞記事内の出現	一致キーワード例
候補者属性	政党情報	政党	カテゴリ	自民党，　政党_立憲民主党，…（※）	◎(自明)	自民党,立憲民主党，日本維新の会
		推薦・支持	カテゴリ	なし／あり		
	プロフィール	経歴	カテゴリ	要職，都道府県議，首長，参院議員，議員秘書，…	○	副代表
		世襲	カテゴリ	世襲かどうか		
		地盤	カテゴリ	地盤かどうか	○	地元，地盤
	選挙区における候補者情報	新旧	カテゴリ	前職,新人,元職	○	新人
		前回出馬	カテゴリ	前回出馬したかどうか		
		当選回数	数値(2)	2未満／2以上		
		選挙区勝率_候補者（候補者の出馬選挙区の勝率）	数値(0.5)	0.5未満／0.5以上		
政党属性	選挙区における政党情報	選挙区勝率_政党（政党の出馬選挙区の勝率）	数値(0.5)	0.5未満／0.5以上		
	政党支持率	政党支持率	数値(0.01,0.03,0.1,...)	閾値未満／閾値以上		
	政党勢い（※自民、立憲民主、維新の候補者のみに属性付与）	政党支持率トレンド_傾き=[政党名（無党派層含む）]	数値(0)	0未満/0以上	○	無党派層
		政党支持率_最新値_自政党-自民党	数値(0)	0未満／0以上：自民党より政党支持率低い／高い（立憲民主か維新）		
		政党支持率トレンド_傾き_自政党-自民党	数値(0)	0未満／0以上：自民党より政党支持率トレンド下向き／上向き（立憲民主か維新）		
		政党支持率_最新値_自政党-立憲民主党	数値(0)	0未満/0以上：立憲民主党より政党支持率低い／高い（自民か維新）		
		政党支持率トレンド_傾き_自政党-立憲民主党	数値(0)	0未満/0以上：立憲民主党より政党支持率トレンド下向き／上向き（自民か維新）	○	維新の風，勢い，突風
		直近都議選で1位政党かどうか	カテゴリ	×		
		直近_都議選勢力（政党割合）	数値(10,40)	閾値未満／閾値以上		
		直近_衆参院再選挙勝率	数値	閾値未満／閾値以上		
地域属性	地理情報	比例 ブロック	カテゴリ	北海道ブロック，東北ブロック，…		
		都道府県	カテゴリ	北海道，…（東京は地区ブロックと重複するため除外）	◎(自明)	大阪
		1区	カテゴリ	YES／NO（1区／1区以外）	◎(自明)	10区
	出馬候補者情報	選挙区候補者数	数値(3)	2人以下／3人以上	○	3人で争う構図,3つどもえの構図
		選挙区に立憲民主党候補含むか	カテゴリ	含まない／含む	○	立憲民主党,立憲，立民
		対決構図	カテゴリ	対決_与野党1対1，対決_与党×野党×野党	○	3人で争う構図,3つどもえの構図,野党共闘
	選挙区・都道府県人口情報	世帯数	カテゴリ	極端に多い/少ない		
		出生者数	カテゴリ	極端に多い/少ない		
		死亡数	カテゴリ	極端に多い/少ない		
		人口数	カテゴリ	極端に多い/少ない		
		出生比率	カテゴリ	極端に多い/少ない		
		人口比率_10代未満（10代,20代，…）	カテゴリ	極端に多い/少ない		
		{団魂，しらけ，…}の世代人口数	カテゴリ	極端に多い/少ない		
		{団魂，しらけ，…}の世代人口比率	カテゴリ	極端に多い/少ない		

※政党名に関しては、学習時・予測時には名寄せをしている

ト上の記事で地盤とされている候補者の他，知事などその地方での政治経験がある候補者を地盤と定義した．選挙区における勝率に関しては，小選挙区制における政党・候補者の勝率を計算した．政党勢いに関しては，政党支持率からトレンド分析を行い，傾きと最新値を比較することで算出した．地理情報については，総務省のサイトから，選挙年の自治体の人口情報を収集し，その情報を選挙区／都道府県に当てはめ，統計分析を行った．「団塊」「しらけ」など世代別の統計情報については，生まれ年代をもとに世代別の人口の概算を算出した．

なお，数値データに関しては，説明の際に閾値が複数あると説明性が低下することを考慮して，1 つの数値属性につき閾値を 1 つとすることを基本とし（政党支持率などを除く），ある数値未満／以上という 2 値化処理を行った．例えば，選挙区における勝率の閾値を 0.1,0.2, …と細かく設定したときに，予測理由に「勝率＜0.2」「勝率＜0.3」の両方が含まれた場合，予測精度が多少向上したとしても，有権者に向けてわかりやすく報道するのが困難になってしまう．そこで，勝率の閾値を 0.5 とすることで「勝率がいい」「勝率が悪い」と説明できるようにした．政党支持率に関しては，選挙の年ごとに変化する値であるため，閾値を１つではなく，複数設定している．当選回数については，新人の当選回数は必然的に0になるため，「新旧_新人」と同等になってしまうことや，ある程度当選している候補者は当選確率が上がるという仮定をおいて，閾値を 2 とした．

3.2 使用する当落予測モデル

2 章で述べた説明変数の組み合わせを線形回帰により予測するモデルとして岩下ら[31]のルールマイニング手法をベースとしたルール線形モデルを用いる．岩下ら[31]のルールマイニング手法では，当落予測などの分類問題の特徴量として説明変数の論理積をデータから高速かつ網羅的に抽出する．その結果得られたルールを用いて，それらの重み付き和をスコア関数とした線形分類モデル（ロジスティック回帰）で学習・予測を行う．特徴量は「変数 A かつ 変数 D かつ 変数 E」のように表されるため，Rusmawati と同様，当落理由は説明変数の組み合わせとみなせる．また，本手法で出力するすべての当落理由には，通常の線形分類モデルと同様に重みが付与される．

例えば，各候補者の説明変数を，属性_属性値のペアで表すことにすると，例えば議員経験の有無に関して，新旧_前職，新旧_元職，新旧_新人などの説明変数を用意できるが，これらは単独（組み合わせ数 d=1）で当落理由となりえる．一方，大阪府では日本維新の会が強い（都道府県_大阪 ∧ 政党_日本維新の会（「∧」は「かつ」と読む））という当落理由は，d=2 のケースに相当する．ルール線形モデルにおいては，説明変数を組み合わせた当落理由を網羅的に探索し，それらを特徴量として線形回帰を行うことにより，当落理由の重みを計算する．本研究においては，対象選挙区において候補者全員の理由を出力し，理由の重みの絶対値を重要度として算出する．その理由から d 個の説明変数の組み合わせである理由を，重要度順に上位 m 件抽出する．Fig.1 に，本手法が入力データから当落予測モデルを構築し，理由のリスト（m=3,d=2 のケース）を抽出するイメージを示す．

例えば 2021 年衆議院選挙の大阪 10 区で，ルール線形モデルは日本維新の会の候補者が当選すると予測しているが，その理由の抜粋を Table 3 に示す．ラベル「TRUE」が当選側の理由，「FALSE」が落選側の理由を示し，重みは，線形回帰で得られた重みを示している．線形回帰モデルのため，当選側の理由には正の重みが，落選側の理由には負の重みがつ

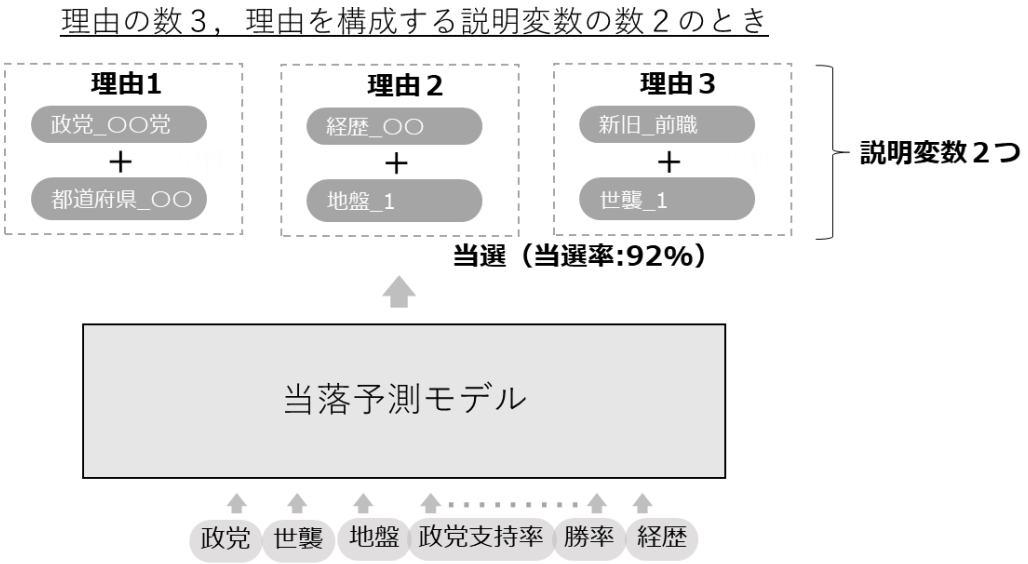

Fig. 1：ルール線形モデルの出力イメージ

Table 3: 理由の例

ラベル	理由	説明変数の数	重み	重要度
TRUE	都道府県_大阪 ∧ 政党支持率トレンド_傾き=自政党-自民党_0未満 ∧ 政党支持率トレンド_傾き=自政党-立憲民主党_0以上	3	3.96	3.96
TRUE	都道府県_大阪 ∧ 政党_日本維新の会 ∧ 政党支持率トレンド_傾き=支持政党なし_0未満	3	0.06	0.06
FALSE	新旧_新人	1	-1.06	1.06

く．「新旧_新人」を含む落選側の理由全ての重みの総和が-3.00 であるものの，「維新の風」に相当する当選側の理由（ラベル TRUE の最初の理由：大阪で，政党勢いが自民党より小さく，立憲民主党より大きい＝大阪で日本維新の会，ラベル TRUE の 2 番目の理由：大阪で，日本維新の会で，無党派層が減少）の全ての重みの総和が+4.01 となっており，モデルの切片も含めた候補者の当選確率は 0.448 と計算されている．これは同じ選挙区の他の候補者の当選確率を上回っており，この候補者が当選と判定している．なお，線形回帰への入力には多数の理由が入力されるが，重みが 0 のものは当落判定には無用のものであるため，実験対象から除外した．

衆議院選挙区の小選挙区制の場合，得票率が最も多い候補者が当選する．本モデルを使い，候補者の属性から当落予想モデルを構築し，各候補者の当選確率を算出する．各選挙区において，最も当選確率が高い候補者 1 人を当選とする．対象選挙区の候補者についての理由を全て含めたものでその選挙区の報道と比較できると考え，重みの絶対値で算出した重要度順に出力する理由の数 m と説明変数の数 d により，理由の有用性の変化を定量的に検証した．

なお，ルール線形モデルにおける 2021 年小選挙区の当落予測の F スコアは，0.747 であった．

3.3 対象選挙区

2021 年衆議院選挙で激戦区だった選挙区を対象とする．特に，複数の新聞記事で「維新の躍進」「大物が落選」など報道されており，その観点からも大阪 10 区を対象とした．

3.4 有用性の評価方法

本研究では，当落予測モデルに基づいて提示する情報とその予測理由の有用性の関係を明らかにすることを目的とする．選挙の専門家が実際に報道した内容を一定レベルでの有用な情報であると仮定し，機械学習から提示された情報を評価する．

3.4.1 評価軸

2 章で述べたように，本論文の研究課題は，予測モデルから出力される理由の数や理由を構成する説明変数の組み合わせの変化がその有用性に与える影響を明らかにすることである．そこで，本論文では，予測モデルが出力する理由の数と 1 つの理由を構成する説明変数の数の 2 軸で，それぞれの値を変化させる．

軸 1：出力する理由の数 m：1～出力数
軸 2：1 つの理由に用いる説明変数の数 d（理由の複雑さ）：1,2,3

この 2 軸の値の組ごとに理由を出力し，新聞記事から抽出したキーワード（3.4.2 節）と一致させる（3.4.3 節）ことにより，どの程度有用な情報を提示できているのかを定量評価する．ここで，有用な情報とは，当該選挙区について書かれている記事内に含まれるキーワードと定義する．

軸 1 に関しては，予測モデルが理由を重要度順に出力するとき，上位 m 件を精査して報道に用いるかどうかを人間が判断することにすると，それがどれくらい有用であるかを示す．軸 2 に関しては，1 つの理由に用いる説明変数の数 d が，その有用性に影響するかどうかを示す．

3.4.2 新聞記事からのキーワードの抽出

対象選挙区に関して，選挙後 1 か月間（2021/10/31～2021/11/30）の間に候補者名で新聞報道（朝日新聞（朝日新聞クロスサーチ[1]），日本経済新聞（日経テレコン[2]））を網羅的に検索し，検索された新聞記事中に含まれる全ての単語から，以下の基準に含まれる単語を除外したものを，キーワードとして抽出した．

- 明示的な単語：選挙，衆院選，衆院議員，など
- 人名：候補者名や議員の名前など
- 選挙「結果」を表す単語：勝つ，破る，敗北，結果，責任，敗因，選挙結果，など
- 日付や時間：10 月 30 日，1 日，夜，未明，早々に，日，など
- 当落理由に関連ない単語：語る，厳しい，前，振り返る，務める，問う，武器，報じる，報道，記者団，関係者，表情，一方，本音，最初，朝日新聞，本当は，持ちこたえる，新たな，難しさ，など

[1] https://xsearch.asahi.com/

[2] http://t21.nikkei.co.jp/

Table 4: 抽出キーワードとその出現頻度

キーワード	出現頻度	キーワード	出現頻度	キーワード	出現頻度
維新	17	***3人で争う構図***	2	府議	1
大阪	14	政権与党	2	代表代行	1
10区	11	***勢い***	2	***与野党対決***	1
前職	8	***共闘***	2	***自民党***	1
自民	7	***党副代表***	2	一騎打ち	1
立民	6	***立憲民主党***	2	***突風***	1
副代表	5	***無党派層***	2	維新支持層	1
新人	5	地盤	2	県内	1
立憲	4	前大阪府議	1	祖父	1
野党共闘	4	***三つどもえの構図***	1	自公政権	1
維新の風	3	***初挑戦***	1	高槻市議	1
府内	2	***大阪府***	1	近畿ブロック	1
自民支持層	2	岸田政権	1	***日本維新の会***	1
新顔	2	対立構造	1	地元	1
政府	2	大阪府知事	1		

※***斜体太字***のキーワードは、説明変数に一致したもの（個数：24）

これらの基準でキーワードを抽出した結果，キーワード種類数は 44，キーワード出現総数は 129 であった（Table 4）．

3.4.3 一致の定義

ルール線形モデルで学習した予測モデルから，対象選挙区における候補者 3 人の予測理由を出力し，そこに含まれる説明変数を抽出した．説明変数の数は 38 であった．38 個の説明変数に 44 種類のキーワードが一致するかどうかのテーブルを作成した．例えば，説明変数「都道府県_大阪」に対しては一致するキーワードは「大阪」「大阪府」が一致する．この例のように自動的に判定できるものの他，以下 3 つの説明変数においては，自動的な判定は難しいが，意味的に一致しているため，それぞれに示すキーワードに一致するとした．

- 「当選_2 未満」：「新人」「新顔」「初挑戦」
- 「1 区_N0」：「10 区」
- 「政党支持率トレンド_傾き=自政党-立憲民主党_0 以上」：「維新の風」「勢い」「突風」

使用属性とそれに一致するキーワードの対応を Table 2 に，理由を構成する説明変数に一致するキーワードを斜体太字で Table 4 に示す．Table 4 から，出現頻度が多いキーワードほど，説明変数に一致していることがわかる．

3.4.4 有用性指標

1 章で述べたように，有権者にとって有用である情報は，いくつかの側面を持つ．新聞記事中に記載される情報は有用だと考えられるが，その中には，有権者にとって既知の「新人」「前職」などの情報も含まれる．一方で，既知の情報「世襲の候補者は強い」を，他の情報と合わせて覆すような情報（「世襲＋〇〇」は弱い）も有用であると考えられる．また記事中の表層的なキーワードだけでなく，記事の意味情報も有用だと考えられる．

本論文では，新聞記事中に記載される情報は有用であるという考えに基づき，定量評価の観点から，政治学の研究で予測に効果的と考えられる属性を使って機械学習で学習したときの予測モデルの理由の有用性を，選挙後に報道された特定選挙区の新聞記事のキーワードとの一致数で定義することを提案する．重要なキーワードほど出現頻度が多いという仮定を置くと，各キーワードの重要性はその出現頻度であると言える．理由がいかに有用かは，その理由を構成する説明変数に一致するキーワードの出現頻度の合計だと考えられる．そこで，有用性を表す指標として，新聞キーワード一致総数 NumKeywordOcc(m,d)を以下のように定義する．

NumKeywordOcc(m,d)＝理由集合において，説明変数の数が d であるような理由の重要度上位 m 件の理由を構成する説明変数に一致するキーワード出現総数/d

この指標は，理由を構成する説明変数 1 つ当たりでどれだけのキーワードに一致したかの総数を表す．一致するキーワードの出現頻度が高いほど，この指標は大きい数値になる．

Table 5: d=2 の出力理由一覧

m	出力理由の候補者政党	理由	m位までの一致キーワード	一致キーワードの出現数合計
1	日本維新の会	選挙区勝率_政党_0.5未満 ∧ 選挙区勝率_候補者_0.5未満		0
2	自民党	選挙区勝率_政党_0.5未満 ∧ 選挙区勝率_候補者_0.5未満		0
3	日本維新の会	直近衆参院再選挙_勝率<0.4 ∧ 選挙区勝率_政党_0.5未満		0
4	自民党	直近衆参院再選挙_勝率<0.4 ∧ 選挙区勝率_政党_0.5未満		0
5	日本維新の会	政党支持率トレンド_傾き=自政党-立憲民主党_0以上 ∧ 直近_都議選勢力<40	勢い,突風,維新の風	6
6	自民党	政党支持率トレンド_傾き=自政党-立憲民主党_0以上 ∧ 直近_都議選勢力<40	勢い,突風,維新の風	6
7	日本維新の会	当選_2未満 ∧ 直近衆参院再選挙_勝率<0.4	初挑戦,勢い,新人,新顔,突風,維新の風	14
8	自民党	当選_2未満 ∧ 直近衆参院再選挙_勝率<0.4	初挑戦,勢い,新人,新顔,突風,維新の風	14
9	日本維新の会	直近衆参院再選挙_勝率<0.4 ∧ 選挙区勝率_候補者_0.5未満	初挑戦,勢い,新人,新顔,突風,維新の風	14
10	自民党	直近衆参院再選挙_勝率<0.4 ∧ 選挙区勝率_候補者_0.5未満	初挑戦,勢い,新人,新顔,突風,維新の風	14
11	日本維新の会	政党支持率トレンド_傾き=自政党-立憲民主党_0以上 ∧ 直近_都議選勢力<20	初挑戦,勢い,新人,新顔,突風,維新の風	14
12	日本維新の会	政党支持率トレンド_傾き=自政党-立憲民主党_0以上 ∧ 直近_都議選勢力<10	初挑戦,勢い,新人,新顔,突風,維新の風	14

4 実験結果

Fig. 2 に，大阪 10 区において新聞キーワード一致総数 NumKeywordOcc(m, d) を計算した結果を示す．

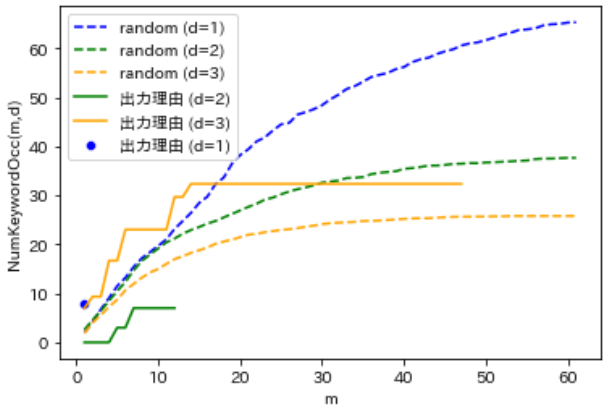

Fig. 2：有用性についての実験結果

まず，m=1 のとき一番有用であるのが，d=1 の出力理由（1 件だけ出力）で，それは「新旧_新人」である．この説明変数に一致するキーワード「新人」「新顔」「初挑戦」の出現頻度合計は 8 である．次に，d=3 のとき，重要度 1 位の理由「都道府県_大阪 ∧ 政党支持率トレンド_傾き=自政党-自民党_0 未満 ∧ 政党支持率トレンド_傾き=自政党-立憲民主党_0 以上」で，「大阪」「大阪府」「勢い」「突風」「維新の風」のキーワードに一致しており，その出現総数は 21 で，NumKeywordOcc(1, 3)＝21/3＝7 である．また，理由数 m が増えるにつれ，新聞キーワード一致総数も上がっており，重要度の高い理由が有用であることが示されている．一方で，d=2 の場合は，重要度上位 4 件の出力理由に一致するキーワードはなかった．重要度 5 位以降で，新聞キーワード一致総数が向上している．

Fig. 2 には，新聞キーワード一致総数のベースラインとして，予測モデルが出力した理由を構成する 38 の説明変数から d 個をランダムに選択し，理由を m 回出力した場合の同指標の平均値を破線で示している．d=1 と d=3 の場合，出力理由の新聞キーワード一致総数はベースラインを上回っている．それに対し，d=2 の場合は，ベースラインを下回っているが，d の変化により有用性が大きく変わるとは考えにくい．Table 5 に d=2 の出力理由一覧を示す．d=2 の重要度上位 2 位までの出力理由は「選挙区勝率_政党_0.5 未満 ∧ 選挙区勝率_候補者_0.5 未満」であった．選挙区における政党および候補者の勝率が両方とも 0.5 未満である，という理由である．この理由に当てはまる候補者は 2 人だった．新聞記事中には，勝率に関するキーワードがなく，一致していなかった．線形ルールモデルの学習対象の属性の中には，新聞記事中では説明されない属性も含まれている．よって，この理由は，新聞記事中には記載されていないが，有用となり得ると考えられる．重要度 3〜4 位の理由も 2 人の候補者に当てはまり，新聞記事中に記載されていなかった．

以上の実験から，予測モデルが出力した理由のうち，d=1 と d=3 の場合に出力理由が有用であることを確認できた．最後に，d=2 の重要度 1〜4 位にあるような理由でこれまで報道されていないような新しい説明ができるのではないか，という観点について検討する．ここでの目的は，我々の手法による出力理由の中で，従来の報道で用いられていない説明が存在する可能性を示すことである．組み合わせ数 d を変化させ，従来は報道されていないような理由の割合がどの程度なのか調査する．

このような新規の理由の可能性を表す指標として，新規理由率 RatioNewReason(m, d) を以下のように定義する．

$$\mathrm{RatioNewReason}(d) = 1-(\mathrm{NumReasonOcc}(d)/\mathrm{NumReason}(d))$$

ここで，新聞キーワード一致理由数 NumReasonOcc(d)，理由数 NumReason(d) を以下のように定義する．

NumReasonOcc(d)＝理由集合において，説明変数の数が d であり，かつ一致するキーワードが

存在する理由の全数
NumReason(d)＝理由集合において，説明変数の数が d である理由の全数

これは，d(=1,2,3)個の説明変数で構成される理由を，それぞれ全件出力した中で，新聞記事中に含まれない説明変数を含む理由の割合を表す．これにより，説明変数の数を変化させることで，新聞が提示出来ていなかった新規の理由を提示できる可能性がどう変化するかを定量評価することができる．38個の説明変数のうち，キーワードに一致しない説明変数の数は11であったが，同じ条件でランダムに説明変数を選択したときの新規理由率と比較した．

新規理由率RatioNewReason(d)の調査結果をTable 6，Fig.3に示す．

Table 6: 新規理由率の定量評価結果

d	NumReason(d)	新規理由率（機械学習出力）	新規理由率（ランダム出力）
1	1	0	0.69
2	12	0.5	0.44
3	47	0.38	0.32

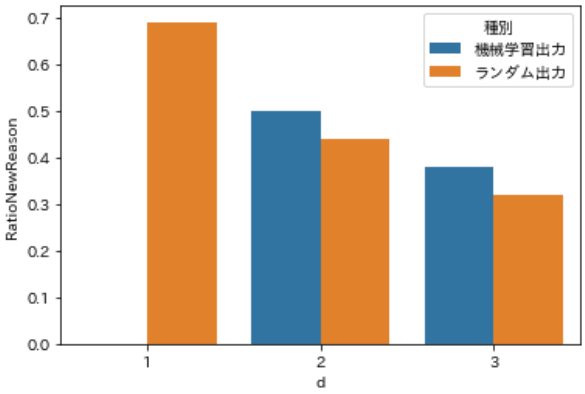

Fig.3: 新規理由率の定量評価結果（グラフ）

この結果によると，説明変数の数が d=2,3 のとき，ランダムに説明変数を選択する場合に比べ，機械学習により出力した理由の方が，新規理由率が高く有効であることがわかった．一方で，d=1 のときは，NumReason(1)＝1 であるため，出力理由にキーワードが一致するか，しないかで新規理由率がそれぞれ 0,1 と大きく変わってしまう．d=1 の場合については今後，対象選挙区を増やして検証していく必要がある．

5 分析・考察

本稿では，選挙の当落予測モデルが出力する予測理由の有用性に焦点をあてた．従来研究では，線形回帰モデルを用いた単一の説明変数による理由や決定木モデルを用いた説明変数の組み合わせによる理由の一例が示されているが，どのような出力形式が有権者にとって有用な情報となるのかは明らかになっていなかった．本論文では，これらの従来研究のモデルの両要素を表現可能なルール線形モデルを採用し，理由の数と各理由を構成する説明変数の数を同時に変化させることで，情報の有用性に与える影響を分析した．その結果，重要な理由ほど有用な情報を提示できている傾向を明らかにした．一方で，政治学で研究されているが新聞記事に含まれないような新規の理由を提示できている可能性を示した．d=2，d=3 の場合，機械学習により出力した理由の新規理由率はランダムで出力した新規理由率より高いので，d や m を複数にして機械学習で理由を出力することには価値があると考える．これにより，メディアが機械学習から出力された理由を報道に使用する際の一つの指針を与えることができ，今後の機械学習の積極的な活用が期待される．

一方で，本論文ではメディアが提示した新聞記事が有用である，という仮定をおいて実験を行っており，今後は有権者の視点での有用性を明らかにする必要がある．具体的には，本論文と同じく理由の数や理由を構成する説明変数の数を変えて，機械学習から出力された理由を報道を想定した情報として構成し，有権者に提示する被験者実験を行う．これにより，有権者の視点から有用な情報提示方法について明らかにする．

参考文献

1) 細貝亮：新聞・テレビはどう伝えたか 第25 回参院選の世論調査報道から，日本世論調査協会報 「よろん」， **124** ，2/12（2019）
2) 松本正生：選挙予測報道の岐路 情勢調査をめぐって，日本世論調査協会報 「よろん」，**125**，26/32（2020）
3) 白崎護：マスメディアに対する選択的接触：2019 年参議院選挙の分析，政策と調査，**18**，43/64（2020）
4) 村上信高，安田英史：『乱！ 総選挙 2012』の番組制作について，映像情報メディア学会誌，**67**-3，263-266（2013）
5) 大坪寛子：政治報道に対する批判，慶応義塾大学メディア・コミュニケーション研究所紀要，**72**，69/83（2022）
6) 放送倫理検証委員会：２０１６年の選挙をめぐるテレビ放送についての意見，放送倫理検証委員会決定第２５号（2017）
7) 山田健太：選挙報道のお行儀～選挙報道・特番は視聴者に何を伝えようとしたのか，民放online（2021）https://minpo.online/article/post-47.html（参照 2023-01-29）
8) Hummel, P., & Rothschild, D.:Fundamental models for forecasting elections. ResearchDMR. com/HummelRothschild_FundamentalModel(2013)
9) Rusmawati, Y.:Automated Reasoning on

Machine Learning Model of Legislative Election Prediction, The 10th International Joint Conference on Knowledge Graphs, 200/204 (2021)
10) Wang, W., Rothschild, D., Goel, S., & Gelman, A.: Forecasting elections with non-representative polls, International Journal of Forecasting, **31**-3, 980/991 (2015)
11) Linzer, D. A.: Dynamic Bayesian forecasting of presidential elections in the states, Journal of the American Statistical Association, **108**-501, 124/134 (2013)
12) 松尾豊：ウェブからの実世界の観測と予測，『電子情報通信学会論文誌 B』，**96**-12，1309/1315（2013）
13) 那須野薫，奥山晶二郎，中西鏡子，松尾豊：Twitter における候補者の選挙地盤に着目した国政選挙の当選者予測，情報処理学会論文誌，**56**-10，2044/2053（2015）
14) 富士通プレスリリース：TBS が「選挙の日 2021」当落速報で，富士通の「説明可能な AI」を活用 ， https://pr.fujitsu.com/jp/news/2021/10/25.html（2021）（参照 2023-01-29）
15) 朝日新聞デジタル：選挙報道，有権者に応えたか 朝日新聞あすへの報道審議会（2021）https://www.asahi.com/articles/DA3S15120002.html（参照 2023-01-29）
16) 飯田健：リスク受容的有権者がもたらす政治的帰結 2012 年総選挙の分析，選挙研究，**29**-2，48/59（2013）
17) 西村翼：政党の公認戦略と地元候補―規定要因としての選挙結果，年報政治学，**71**-2，2_280/2_302（2020）
18) 原田隆司：自民党衆議院議員の経歴パターン分析 昭和五五年総選挙当選者について，ソシオロジ，**29**-1，45/67（1984）
19) 吐合大祐：選挙区定数と議員の再選戦略: 日本の都道府県議会議員の委員会所属に注目して，年報政治学，**69**-1，1_293/1_315（2018）
20) 平野浩：選挙研究における 「業績評価· 経済状況」 の現状と課題，選挙研究，**13**，28/38（1998）
21) 善教将大：政党支持は投票行動を規定するのか―サーベイ実験による長期的党派性の条件付け効果の検証―，年報政治学，**67**-2，2_163/2_184（2016）
22) 寺井晃：選挙制度についての分析: 政党支持率と獲得議席数の乖離についてのシミュレーション，京都産業大学論集. 社会科学系列，**29**，143/154（ 2012）
23) 三宅一郎：選挙制度変革と候補者要因，情報研究: 関西大学総合情報学部紀要，**13**，51/72（2000）
24) 山田真裕：投票率の要因分析，選挙研究，**7**，100-116（1992）
25) 名取良太：2012 年衆院選における政党投票と候補者投票，情報研究: 関西大学総合情報学部紀要，**41**，71/84（2014）
26) 岡田浩：党派性の地域的偏りの要因―「自民党王国」 北陸・石川の検討，金沢法学，**59**-1，27/48（2016）
27) 加藤元宣：小選挙区の地域特性に基づく 2000 年衆院選の分析，選挙研究，**17**，154/170（2002）
28) 品田裕：2009 年総選挙における選挙公約，選挙研究，**26**-2，29/43（2011）
29) 堀内匠：得票分析にみる 2009 年東京都議会議員選挙と衆議院議員総選挙の連続性，自治総研，**35**-9，62/92（2009）
30) 権藤敏範：「衆院２補選と統一地方選 今後の影響は」（時論公論），NHK 解説委員室（ 2019 ） https://www.nhk.or.jp/kaisetsu-blog/100/318927.html（参照 2023-02-22）
31) 岩下洋哲，高木拓也，鈴木浩史，後藤啓介，大堀耕太郎，有村博紀：密なデータベースに対する動的な探索順序を用いた高速な顕在パターンマイニング手法，人工知能学会研究会資料 人工知能基本問題研究会 111 回，40/45（2020）