情報処理学会論文誌　数理モデル化と応用　Vol.14　No.3　39–50 (Aug. 2021)

機械学習による中古車落札価格の要因分析および割安評価

工藤 大輝[1]　福西 亮介[2]　黛 広樹[3]　鈴木 智也[1,a)]

受付日 2020年11月16日，再受付日 2021年1月4日，
採録日 2021年2月5日

概要：本研究では，全国のオートオークションで落札された中古車ビッグデータを対象として，重回帰式ベースのヘドニックアプローチと非線形モデルの XGBoost に基づいて個車価格推定モデルを構築し，正答率と各特徴量の重要度やポジネガ極性について比較分析した．中古車の各特徴量は非線形関係を有していることから XGBoost の正答率が良好であり，重回帰式のみではとらえられない特徴量の重要性を確認した．またオートオークション会場は全国各地にあり，地域や規模に応じて落札価格に「割安」や「割高」などの特徴が生じる可能性がある．そこで非線形な個車価格推定モデルを利用することで割安な落札価格を検出し，裁定取引などへの応用可能性を紹介する．

キーワード：価格査定，機械学習，データサイエンス，AI の説明可能性

Factor Analysis and Undervalued Prices in Used Car Auctions by Machine Learning

Daiki Kudo[1]　Ryosuke Hukunishi[2]　Hiroki Mayuzumi[3]　Tomoya Suzuki[1,a)]

Received: November 16, 2020, Revised: January 4, 2021,
Accepted: February 5, 2021

Abstract: We estimated used-car prices contracted in auto-auction markets by using the hedonic approach based on a linear regression and the XGBoost model based on a nonlinear machine learning, and evaluated their accuracies and the importance and polarity of explanatory variables. As a result, we confirmed some advantages of using machine learning because each used car has unique properties and these features are nonlinearly interacted. Moreover, we applied our price evaluation model to detect undervalued prices as arbitrage opportunities caused by the location and scale of auto-auction markets.

Keywords: price assessment, machine learning, data science, explainable AI

1.　はじめに

中古車はユーザから買取店へ流れた後，主にオートオークションを通じて販売店に売却される．そこで買取店はオートオークションでの落札価格を予想して，ユーザに対し買取価格を提示する必要がある．その際に合理的な買取価格を算出する査定システムが必要となるが，中古車や住宅のような不動産は同一物が存在しないため価格モデルの構築が難しい．そこで伝統的には，重回帰式に基づくヘドニック法 [1], [2], [3], [4], [5] によって不動産を構成する各属性の重要度を推計し，それらの線形結合によって価格モデルを構築する方法が主流である．たとえば，国土交通省が公表する不動産価格指数 [6] は，このヘドニック法をベースに算出している．しかし近年のソフトコンピューティング技術の進歩により，ビッグデータや機械学習を積極的に活用する事例 [7], [8] が増えており，関連学会誌において不動産テックの特集が組まれている（人工知能学会誌，日本機械学会誌，日本不動産学会誌，など）．しかしいずれも住宅販売を対象にしており，中古車への応用可能性について検討されていない．そこで本研究では，全国のオートオー

1　茨城大学大学院理工学研究科
Graduate School of Science and Engineering, Ibaraki University, Hitachi, Ibaraki 316–8511, Japan
2　株式会社 KINTO
KINTO Corporation, Nagoya, Aichi 450–0002, Japan
3　株式会社プロトコーポレーション
PROTO Corporation, Nagoya, Aichi 460–0006, Japan
a)　tomoya.suzuki.lab@vc.ibaraki.ac.jp

クションで落札された中古車の長期実績（約 160 万件）をビッグデータと見なし，これらを機械学習することで価格査定モデルの精度を向上できるか検証する．

機械学習を導入する利点は，各属性が有する非線形性への対応である．たとえば排気量は多いほど価格を上昇させる要因であるが，税金や保険料が高くなるため非線形的に作用する [9]．また年式においても古いほど価格を低下させる要因であるが，あえてビンテージ車を好むユーザ層も存在する．これらが価格に与える影響力においても，メーカや車種のブランドによって複雑に変化すると考えられる．同様に中古住宅においても価格構造の非線形性が報告されている [10]．一方，機械学習を導入する短所として解釈性の低下が考えられるが，各属性の重要度（Feature Importance）を評価できる決定木ベースのアルゴリズムを用い，さらに従来のヘドニック法による偏回帰係数も参照することで，重要度のポジネガ極性を評価する．なお同種の目的として，近年では Explainable AI（XAI）といった機械学習の説明力を高める工夫が注目されている．特に LIME（Local Interpretable Model-agnostic Explanations）[11] や SHAP（SHapley Additive exPlanations）[12] が代表的であり，本研究では SHAP の観点からも各属性の重要度やポジネガ極性を評価する．

機械学習によって価格査定モデルを強化できれば，応用事例の 1 つとして，落札価格の割高や割安について評価できるだろう．価格査定モデルで個車状態に応じた適性価格を算出し，もし実際の落札価格が適性価格より低ければ割安と評価する．オートオークションの落札価格は大人数による合意（集合知）であるため合理性が高いと考えられるが，地方や小規模なオートオークション会場では市場参加者が少ないため，合理性な価格形成が困難かもしれない．もしそうならば，割安な会場で現物を仕入れ，割高会場で転売するといったディーリングも考えられる．そこで価格査定モデルの応用として，落札価格の割高・割安評価を試みる．

2.　分析データ

本研究では，2011 年 1 月から 2017 年 12 月において日本全国のオートオークション会場（全 78 会場）で取引された中古車の落札実績データを使用する．対象車種は利用者の広範な年齢層や用途を考慮し，**表 1** に示す 20 車種とする．

学習器に投入する説明変数を**表 2** に示す．「新車価格」は単に価格を表すだけでなく，同一車種におけるグレードの違いを表現できる．「流通台数」は需要と供給の関係により，流通台数が多くなれば価格を下げる要因になりうる．「排気量」は軽自動車のように 1 種類のみの場合は除外した．質的変数については one-hot 表現によるダミー変数に変換した．

表 1　分析対象車種
Table 1　Car models for analysis.

ボディータイプ	車種	メーカ	落札実績数
セダン	LS	レクサス	34,472
	CROWN	トヨタ	214,160
	PRIUS	トヨタ	371,601
コンパクト	FIT	ホンダ	348,034
	DEMIO	マツダ	124,969
	AQUA	トヨタ	92,362
	NOTE	日産	140,315
	Vitz	トヨタ	390,493
ミニバン	SERENA	日産	211,484
	ESTIMA	トヨタ	108,321
	STEP WGN	ホンダ	191,471
SUV	Forester	SUBARU	75,497
	PAJERO	三菱	40,042
	PRADO	トヨタ	112,209
	OUTLANDER	三菱	26,064
スポーツ	Roadstar	マツダ	16,364
軽自動車	TANTO	ダイハツ	168,783
	WagonR	スズキ	244,976
	ALTO	スズキ	80,064
	MOVE	ダイハツ	299,603

表 2　説明変数
Table 2　Explanatory variables.

量的変数［単位］	
走行距離	出品時の走行距離［1,000 km］
落ち年	出品年と新車登録年の差［年］
評価点	内装，外装の総合状態（0〜5）［点］
車検残月	出品時の車検の残り月数［カ月］
排気量	エンジン排気量［cc］
流通台数	推定月前月の全会場の出品台数［台数］
新車価格	新車時の価格［円］
質的変数（ダミー変数，該当すれば 1 を付与）	
オプション	ナビ&テレビあり，ナビのみ レザーシートあり，サンルーフあり
修復歴	修復歴（事故歴を含む）あり
取引月	1 月，2 月，3 月，4 月，5 月，6 月 7 月，8 月，9 月，10 月，11 月，12 月
車体色	白，真珠，黒，黄，灰，金，シルバー 琥珀，紫，アイボリー，青，青黒 赤，茶，鉄，ピンク，濃灰，濃緑 薄黄，薄緑，葡萄，オレンジ，その他
取引地域	北海道，東北，関東，中部，近畿 中国・四国，九州・沖縄

3.　重回帰式による個車落札価格の推定

ヘドニックアプローチは，財の価格をその財の属性を基に回帰して属性の計算価格を推定し，属性の量と計算価格の積和によって，多様な財の価格を評価する方法論であ

る [3]．本研究においては，過去に落札された中古車価格 y_i とその属性 $x_{p,i}$（$p = 1, 2, \cdots, P$）に基づいて重回帰分析を行い，推定された偏回帰係数 $\hat{\alpha}_p$ を各属性の計算価格と見なす．

$$y_i = \hat{\alpha}_1 x_{1,i} + \hat{\alpha}_2 x_{2,i} + \cdots + \hat{\alpha}_P x_{P,i} + \epsilon_i \qquad (1)$$

ここで ϵ_i は残差であり，i は学習期間における中古車のインデックスである．説明変数 x_p の量的変数については標準化を施し，目的変数 y はそのまま用いる．そして最小 2 乗法により $\hat{\alpha}_p$ を推定する．なお多重共線性への対応として，ダミー変数化した質的変数については基準を設け，基準となるダミー変数は説明変数 x_p から除外する．取引月では「1 月」，車体色では「シルバー」，取引地域では「関東」を基準とする．この処置は主に説明モデルとして用いられる重回帰式では一般的であるが，後述する機械学習では汎化精度を優先するため，すべてのダミー変数を説明変数 x_p として用いる．

次に，学習期間以降に出品される中古車のインデックスを j と標記すると，次式のように落札価格を推定できる．

$$\hat{y}_j = \hat{\alpha}_1 x_{1,j} + \hat{\alpha}_2 x_{2,j} + \cdots + \hat{\alpha}_P x_{P,j} \qquad (2)$$

価格推定においても学習期間と同じ標準化を適用するが，$\hat{y}$ については標準化の逆変換によって元の価格スケール［円］に戻す．なお本章では，2017 年 1 月から 2017 年 12 月を価格推定期間とし，1 カ月ごとに推定を行う．その際に，推定月前月から過去 2 年間を学習期間とする．なお，表 1 に示す車種別に学習（式 (1)）および推定（式 (2)）を行う．

式 (1) および式 (2) の前処理として，説明変数間の多重共線性について配慮する．本研究では，Python の statsmodel ライブラリに搭載されている VIF（分散拡大係数）を計算し，VIF が 10 を超える説明変数 x_p は多重共線性の原因となる可能性が高いため，各月の重回帰式から除外した．その際に，最大の VIF を示す説明変数を 1 つ除外した後に VIF を再計算し，残された説明変数の VIF がすべて 10 未満になるまで 1 つずつ除外した．残された説明変数は車種ごとおよび月ごとで異なるため，具体的な記載は割愛する．

データへのあてはめや偏回帰係数の解釈の便宜性から，ヘドニックアプローチでは目的変数のみに対数変換を施す半対数型を用いることが多い [3]．たとえば，目的変数である住宅価格のみに対数変換を施し，階数や建築年数といった説明変数は無変換で重回帰式に適用している [13]．そこで本研究においても説明変数は無変換とし，目的変数について対数変換と無変換の 2 通りで推定を行う．この理由として，5 章で導入する非線形モデルでは目的変数の対数変換を必要としないため，本章の線形モデルと比較可能にするためである．対数変換を適用する場合は，式 (1) では y_i を $\log y_i$ とし，式 (2) では $\hat{y}_j$ を $\log \hat{y}_j$ とする．その後，$\exp(\log \hat{y}_j)$ によって $\hat{y}_j$ を得る．

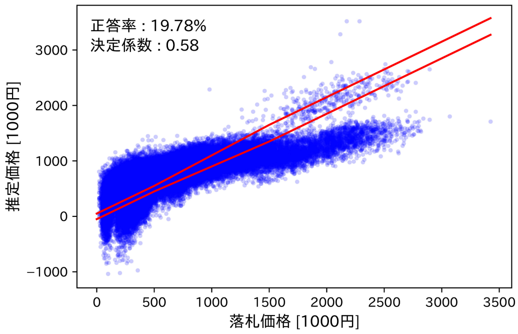

図 1　推定価格 $\hat{y}_j$（縦軸）と真の落札価格 y_j（横軸）の相関図．ただし目的変数に対数変換を施さない場合．図中の各点は落札された PRIUS の個車であり，上下の実線は誤差の許容範囲である

Fig. 1　Correlation diagram between the estimated price $\hat{y}_j$ shown in the vertical axis and its actual price y_j shown in the horizontal axis without applying the logarithmic transformation to the objective variable. Each dot corresponds to individual car of PRIUS, and two solid lines show the upper and lower limits of the acceptable error range.

推定価格 $\hat{y}_j$ の精度評価として，式 (2) における決定係数 R^2 と正答率 A を算出する．決定係数 R^2 は 1 に近いほど学習期間以後の中古車 j に対する価格推定モデルのあてはめ精度が高く，正答率 A も同様である．ただし正答率 A を算出する際には実務利用を考慮し，以下の条件を満たす場合に「正答」と見なした．

- if $y_j \leq 50$ 万円，$y_j - 5$ 万円 $\leq \hat{y}_j \leq y_j + 5$ 万円
- if 50 万円 $< y_j <$ 150 万円，$0.9 \cdot y_j \leq \hat{y}_j \leq 1.1 \cdot y_j$
- if 150 万円 $\leq y_j$，$y_j - 15$ 万円 $\leq \hat{y}_j \leq y_j + 15$ 万円

ここで y_j は実際の落札価格であり，$\hat{y}_j$ はその推定値である．この条件の意図として，50 万円〜150 万円がオートオークションにおける落札価格のボリュームゾーンであり，この範囲については ±10%以内の誤差を許容する．しかし 150 万円以上の高価格帯においては許容誤差が青天井に拡大するため，±15 万円を許容誤差の範囲とする．一方，50 万円以下の低価格帯においては許容誤差が 0 まで縮小するため，±5 万円を許容誤差の範囲とする．最後に，正答数を推定台数で割ることで正答率 A を算出する．

式 (2) に基づいて落札価格を推定した様子を**図 1**，**図 2** に示す．なお代表的な事例として，PRIUS の結果を示す．対数変換の有無を比較すると，図 1 の対数変換なしの場合は高価格帯では過小評価される個車が多く，全体的に分布が曲がっている．一方，図 2 のように対数変換を施すことにより，過小評価は改善され，より直線的な分布形状を示している．

次に全対象車種について，正答率 A と決定係数 R^2 を集計した結果を**表 3** に示す．やはりほとんどの車種において，対数変換を施した方が推定精度が高い．しかし一部の軽自動車や SUV において推定精度が低下する場合もある．

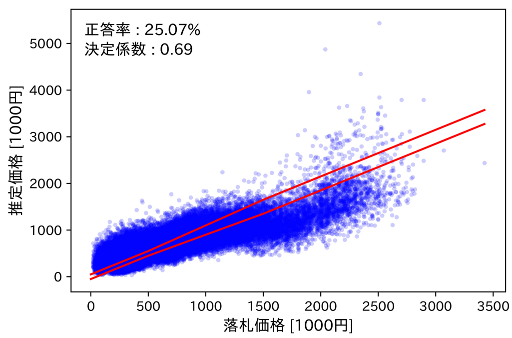

図 2　図 1 と同様．ただし目的変数に対数変換を施した場合
Fig. 2　Same as Fig. 1, but applying the logarithmic transformation to the objective variable.

表 3　推定価格 $\hat{y}_j$ の正答率 A と決定係数 R^2
Table 3　The accuracy A and the coefficient of determination R^2 of the estimated price $\hat{y}_j$.

車種	正答率 A [%]		決定係数 R^2	
	無変換	対数変換	無変換	対数変換
LS	4.91	7.48	0.45	0.57
CROWN	10.93	18.72	0.44	0.53
PRIUS	19.78	25.07	0.58	0.69
SERENA	13.95	23.43	0.62	0.58
ESTIMA	10.10	14.36	0.50	0.66
STEP WGN	16.07	40.60	0.46	0.48
TANTO	25.36	32.31	0.74	0.80
WagonR	51.85	50.99	0.89	0.86
ALTO	49.70	54.78	0.86	0.84
MOVE	44.87	51.87	0.87	0.86
FIT	18.70	38.03	0.40	0.46
DEMIO	34.31	60.24	0.72	0.86
AQUA	20.10	21.21	0.33	0.41
NOTE	22.14	26.10	0.60	0.66
Vitz	23.42	49.94	0.71	0.83
Forester	25.98	41.68	0.88	0.66
PAJERO	22.05	37.37	0.84	0.83
PRADO	27.31	24.66	0.90	0.82
OUTLANDER	12.98	17.11	0.37	0.54
Roadstar	20.11	23.12	0.73	0.80

4.　落札価格に影響する各属性の非線形性

重回帰式による個車価格推定において，対数変換を施すことにより各評価指標は向上したが，決定係数 R^2 は 0.4〜0.8 程度にとどまっている．この原因の 1 つとして，重回帰式では表現できない非線形構造の存在が考えられる．なお，中古住宅においても価格形成メカニズムに非線形性の存在が報告されており [10]，中古車においても例外ではないと示唆される．そこで，学習期間（2015 年 1 月〜2016 年 12 月）に落札された個車の落札価格 y_i と各属性 $x_{p,i}$ の関係性を調べ，式 (1) のようなシンプルな重回帰式では表現できない非線形性を確認した．その代表的な事例を紹介する．

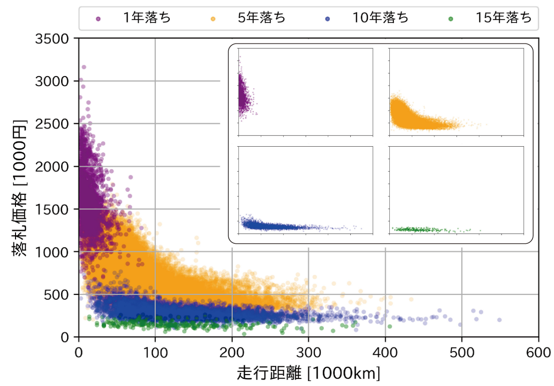

図 3　走行距離と落札価格の関係（PRIUS の例）
Fig. 3　Relationship between mileages and contract prices of PRIUS.

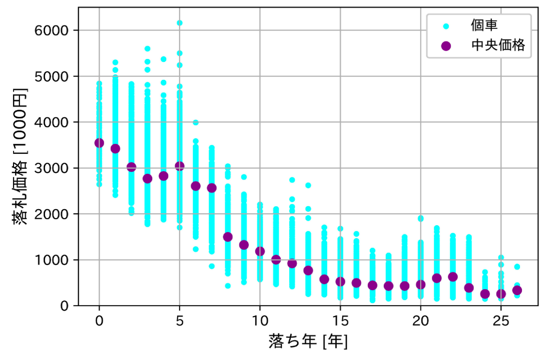

図 4　落ち年と落札価格の関係（PRADO の例）
Fig. 4　Relationship between car ages and contract prices of PRADO.

図 3 は，PRIUS の落ち年別の走行距離に対する落札価格の分布である．1 年落ちや 5 年落ちのように新しい個車においては走行距離は大きな減価要因であるが，10 年落ちや 15 年落ちのように古い個車ではその感応度は低下し，走行距離による減価率は小さい．つまり感応度が落ち年の影響を受けるため，走行距離と落ち年によるクロスターム項の導入が考えられる．しかし他の属性の影響も考慮すると膨大な組合せが考えられるので，モデルの解釈性を最優先にしないならば，非線形な機械学習によって機械的に説明変数を設計することもできる．この点については，次章で検討する．

図 4 は，PRADO の落ち年に対する落札価格の分布である．落ち年が大きくなるほど，中央価格（個車の中央値）減衰は緩やかになる．さらに，8 年落ちにおける不連続的な下落や，22 年落ちにおける局所的な上昇は，おそらくモデルチェンジが関係している．前者については前モデルの需要が減り，後者については前々モデルがむしろビンテージ車として増価したと推察される．

以上のように，中古車の属性には落札価格に対して非線形的に影響する可能性がある．しかしその可能性は多様に考えられるため，次章において機械学習を導入し，有用な非線形モデルの自動構築を試みる．

5. 非線形モデルによる個車価格推定

前章で確認したように，各属性が落札価格に与える影響は必ずしも重回帰式で表現できるとは限らない．そこで機械学習を導入することで，効果的な非線形モデルを機械的に探索する．しかし機械学習によって最適化されたモデルは重回帰式に比べて解釈性が低下するため，比較的解釈しやすい決定木ベースの機械学習を導入する．これにより説明変数の重要度を算出できるため，6 章で述べる要因分析に利用する．さらに集団学習によって決定木の学習性能を高めるべく，勾配ブースティングを組み合わせた XGBoost [14] を使用する．なお XGBoost の概要については，付録 A.1 に記載する．

重回帰式である式 (1)，(2) を XGBoost の関数 F に置き換えると，それぞれ

$$y_i = F(x_{1,i}, x_{2,i}, \cdots, x_{P,i}) + \epsilon_i \quad (3)$$

$$\hat{y}_j = F(x_{1,j}, x_{2,j}, \cdots, x_{P,j}) \quad (4)$$

となる．式 (3) で関数 F を学習し，式 (4) で未学習データに適用する．関数 F は非線形性を表現できるため，目的変数 y_i には対数変換を適用しない．さらに機械学習はモデルの解釈性を重視しないため，多重共線性を考慮せずにダミー変数の基準を含むすべての説明変数を入力する．その他，価格推定期間や学習期間などについては，3 章の重回帰分析と同一とする．

XGBoost の学習において，損失関数 l は 2 乗誤差を使用し，5 分割交差検証法により汎化精度を求める．学習率は 0.03 とし，アーリーストッピングによって汎化精度の改善が停止するまで決定木の個数 K を増加させた．罰則パラメータはデフォルト値（$\lambda = 1$，$\gamma = 0$）を使用し，決定木の深さの上限値と補正量 ω_m の下限値は汎化精度に基づいて最適化した．

図 2 と同様に，PRIUS の落札価格を推定した様子を**図 5** に示す．図 2 の重回帰式（対数変換あり）と比較して，高価格帯においても安定的に落札価格を推定できている．分布形状もより直線的になっていることから，推定誤差の縮小を確認できる．

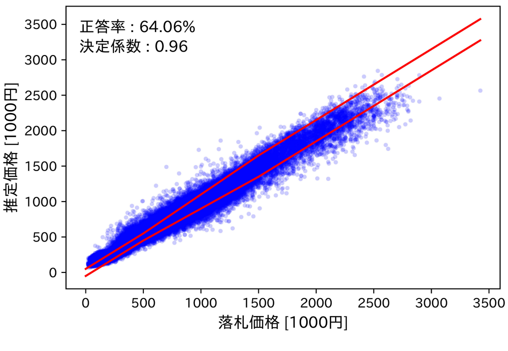

図 5　図 1 と同様．ただし価格推定モデルに XGBoost を適用した場合

Fig. 5　Same as Fig. 1, but using XGBoost to estimate $\hat{y}_j$.

次に全対象車種について，正答率 A と決定係数 R^2 を集計した結果を**表 4** に示す．すべての車種において推定精度が改善されている．特に正答率 A は 20～50%も向上し，軽自動車やコンパクトカーのような大衆車種において 70%以上の精度を実現している．このように XGBoost による機械学習モデルの恩恵を確認できる．

6. 要因分析

学習後の重回帰式（対数変換あり）と XGBoost を観察することで，中古車の各属性がどのように落札価格に寄与するのかを分析する．重回帰式においては式 (1) で推定された偏回帰係数 $\hat{\alpha}_p$ を説明変数 x_p の重要度と見なし，有意水準 5%の t 検定を適用する．さらに $\hat{\alpha}_p$ の符号でポジネガ極性を評価する．なお，目的変数 y に対数変換を施したことにより，$\hat{\alpha}_p = \frac{\mathrm{d} \log y}{\mathrm{d} x_p} = \frac{\mathrm{d}y/y}{\mathrm{d}x_p}$ となるため，偏回帰係数 $\hat{\alpha}_p$ は落札価格 y の変化率に対する説明変数 x_p の寄与を表す．

XGBoost においては式 (A.6) のゲイン G_m に基づき，以下のように説明変数 x_p の重要度 I_p を算出する．

$$I_p = \frac{1}{|M_p|} \sum_{m=1}^{|M_p|} G_m^* \quad (5)$$

ここで XGBoost を構成するすべての決定木において，M_p

表 4　XGBoost による推定価格 $\hat{y}_j$ の正答率 A と決定係数 R^2．カッコ内に表 3 の重回帰式（対数変換）からの改善量を示す

Table 4　The accuracy A and the coefficient of determination R^2 of $\hat{y}_j$ estimated by XGBoost. Each figure in parenthesis shows the improvement from the multiple linear regression with logarithmic transformation shown in Table 3.

車種	正答率 A [%]		決定係数 R^2	
LS	31.43	(+23.87)	0.94	(+0.37)
CROWN	56.85	(+38.13)	0.97	(+0.44)
PRIUS	64.06	(+38.99)	0.96	(+0.27)
SERENA	66.72	(+43.29)	0.98	(+0.39)
ESTIMA	52.53	(+38.17)	0.96	(+0.30)
STEP WGN	69.80	(+29.20)	0.96	(+0.40)
TANTO	69.32	(+37.01)	0.97	(+0.16)
WagonR	75.31	(+24.32)	0.96	(+0.11)
ALTO	74.70	(+19.92)	0.95	(+0.11)
MOVE	75.47	(+23.60)	0.97	(+0.10)
FIT	75.86	(+37.83)	0.95	(+0.49)
DEMIO	74.18	(+13.94)	0.96	(+0.10)
AQUA	72.76	(+51.55)	0.94	(+0.52)
NOTE	72.43	(+46.33)	0.96	(+0.30)
Vitz	73.30	(+23.36)	0.95	(+0.12)
Forester	69.27	(+27.59)	0.98	(+0.32)
PAJERO	60.90	(+23.53)	0.97	(+0.14)
PRADO	51.34	(+26.68)	0.96	(+0.15)
OUTLANDER	53.16	(+36.05)	0.93	(+0.39)
Roadstar	51.87	(+28.75)	0.96	(+0.17)

は説明変数 x_p に基づいて分割されたノード m の集合を表し，G^*_m は実際のノード分割で得られた最大ゲインを表す．このように M_p における平均ゲインにより，説明変数 x_p の重要度 I_p を評価する [15]．この I_p は相対的な評価基準なので，解釈しやすいように $\sum_{p=1}^{P} I_p = 1$ によりスケール調整を施す．しかしポジネガ極性を評価できないため，XGBoost については SHAP [12] の観点からも分析する．なお SHAP の概要を付録 A.2 に示す．

式 (A.8) の SHAP 値 $\phi_p(\boldsymbol{x}_i, F)$ は中古車 i ごとに算出されるため，SHAP による説明変数 x_p の重要度 S_p およびポジネガ極性 r_p を以下のように算出する．

$$S_p = \frac{1}{n}\sum_{i=1}^{n} |\phi_p(\boldsymbol{x}_i, F)| \tag{6}$$

$$r_p = \frac{1}{\sigma_{\phi_p}\sigma_{x_p}} \cdot \frac{1}{n}\sum_{i=1}^{n}(\phi_p(\boldsymbol{x}_i, F) - \bar{\phi}_p)(x_{p,i} - \bar{x}_p) \tag{7}$$

SHAP 値 ϕ_p と説明変数 x_p の相関係数 r_p が正ならば，式 (A.7) および $F(\boldsymbol{x}_i) \simeq g(\boldsymbol{x}_i)$ より，説明変数 x_p はモデル F の出力値に対してポジティブ極性を持ち，相関係数 r_p が負ならばネガティブ極性を持つと解釈できる．なお $\phi_p(\boldsymbol{x}_i, F)$ は $F(\boldsymbol{x}_i)$ と同じ単位（[円]）を持ち，重要度 S_p も [円] の単位を持つためスケール変換は行わない．

分析結果を**図 6**，**図 7**，**図 8**，**図 9**，**図 10** に示す．いずれも 2017 年の各月から過去 2 年間を学習期間とし，重回帰式に関する $\hat{a}_p$ および XGBoost に関する I_p，S_p，r_p を算出した．なお誌面の都合により，車種を厳選して典型的な事例を示す．

図 6 においては，落ち年の重要度が高く，落札価格を下げるネガティブ要因である．具体的には，S_p によれば落ち年の有無によって落札価格は 40 万円程度変化し，$\hat{a}_p$ によれば落ち年の 1σ 増加に対して落札価格は 65%程度低下する．同様に I_p や r_p からも，高い重要度およびネガティブ極性を確認できる．

図 7 においては，レザーシートの評価に揺らぎが見受けられる．偏回帰係数 $\hat{a}_p$ は統計的に有意な影響力を示していないが，XGBoost による S_p や I_p ではレザーシートの影響力は高い．この原因として，レザーシートの有無が評価点に影響し，評価点を通じて落札価格に影響を与えたと推察される．重回帰式では目的変数に対する線形的な直接効果しか評価できないが，XGBoost ならば非線形的な間接効果も含めて総合的に評価できる．しかし S_p と I_p においてもレザーシートとサンルーフの評価に差がある．そこでドメイン知識として，PRADO の上位モデルはレザーシートが標準装備であり，サンルーフはモデルを問わずつねに追加料金が必要となる．つまりサンルーフの方が増加要因になりやすいため，S_p の方が整合的である．文献 [12] で

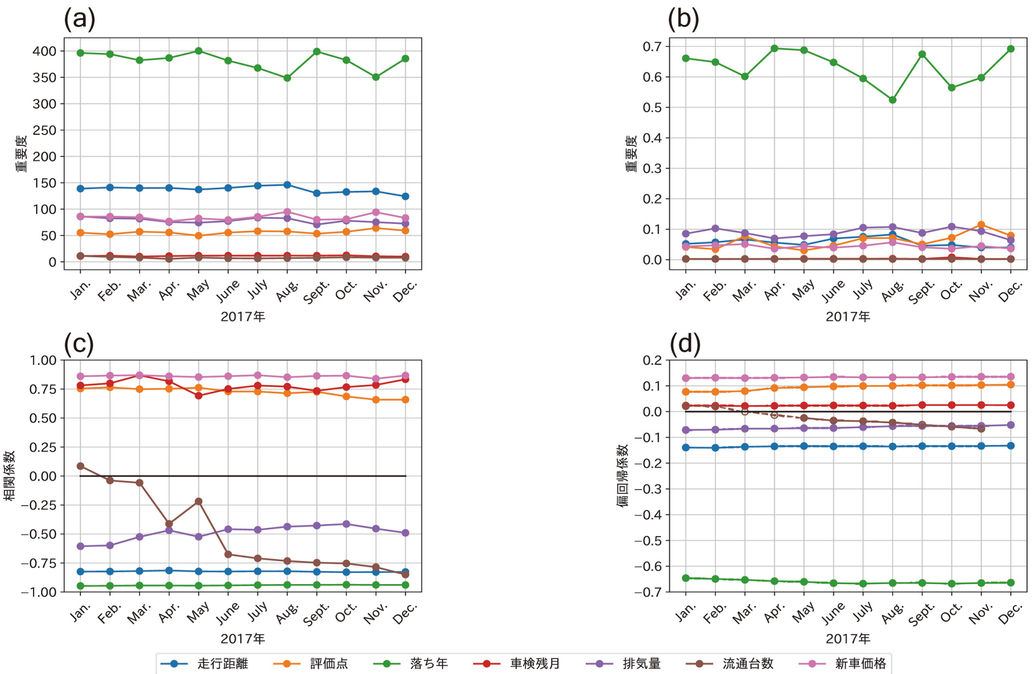

図 6　量的変数の要因分析（PAJERO の例）(a) SHAP 値による重要度 S_p [1,000 円]，(b) ゲインによる重要度 I_p，(c) SHAP 値によるポジネガ極性 r_p，(d) 重回帰式による偏回帰係数 $\hat{a}_p$（点線は有意水準 5%の t 検定において有意でない場合）

Fig. 6　Factor analysis of quantitative variables in PAJERO: (a) the importance S_p [1,000 yen] calculated by SHAP value, (b) the importance I_p estimated by the gain, (c) the polarity r_p calculated by SHAP value, (d) the partial regression coefficient $\hat{a}_p$ estimated by the multiple linear regression where dotted lines mean the cases that $\hat{a}_p$ cannot pass the t-test at the 5% significance level.

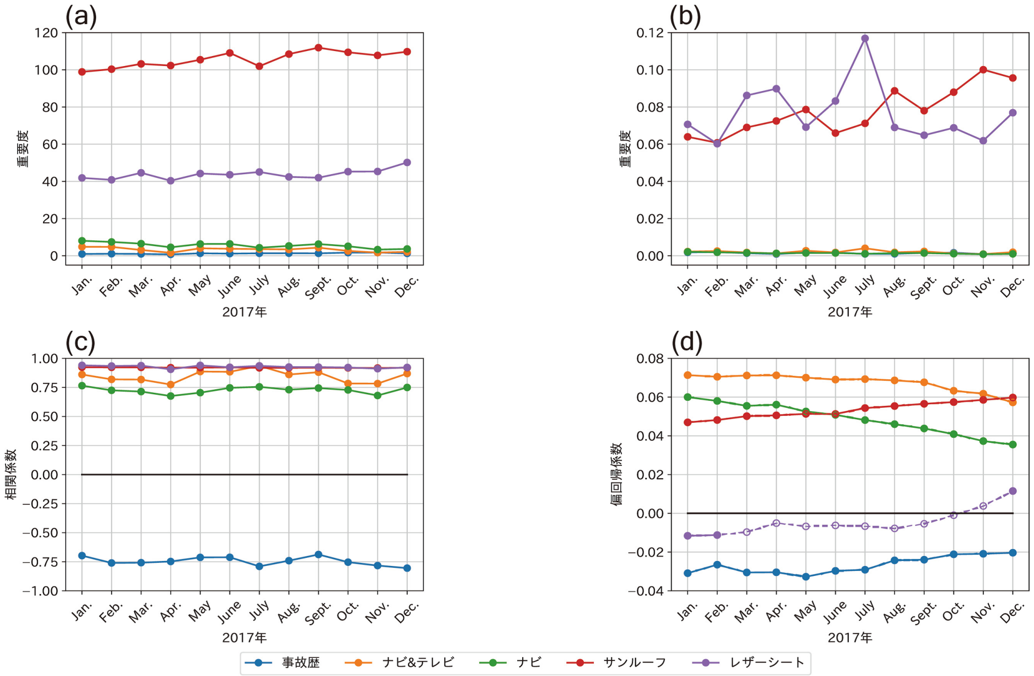

図 7　図 6 と同様，ただしオプション装備の要因分析（PRADO の例）

Fig. 7　Same as Fig. 6, but factor analysis of optional equipments in PRADO.

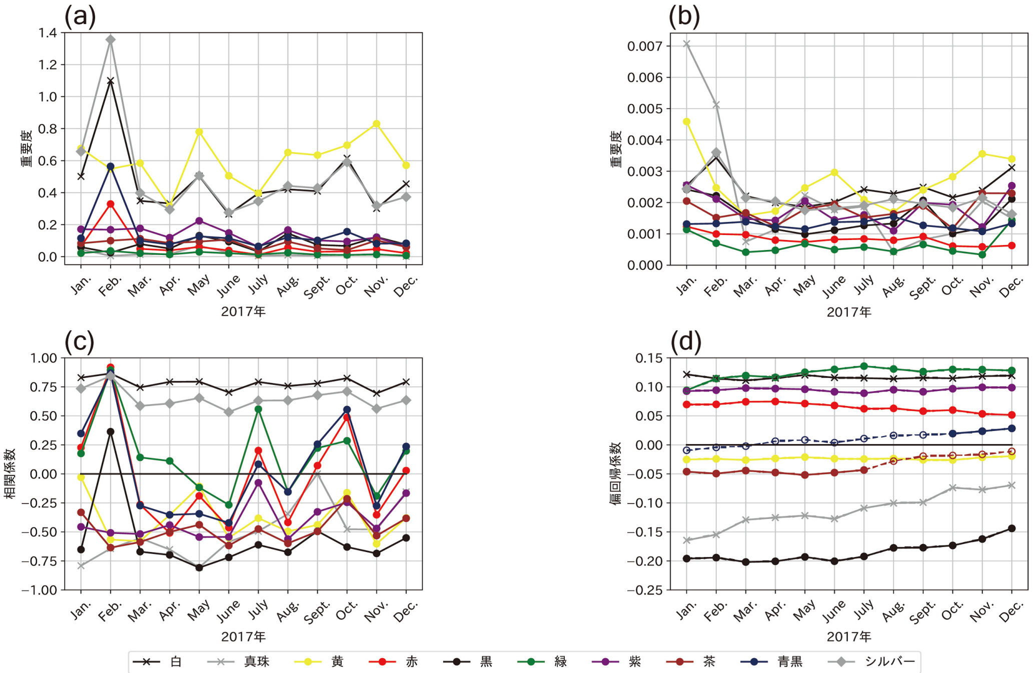

図 8　図 6 と同様，ただし車体色の要因分析（DEMIO の例）

Fig. 8　Same as Fig. 6, but factor analysis of body colors in DEMIO.

も S_p の方が I_p よりも人間の判断と一致しやすい傾向が指摘されており，本分析結果も同傾向を示している．なお，人間の判断のみでは主観的になりやすいが，SHAP による統計分析によって客観性を担保できる．

図 8 においては車体色による影響は全体的に小さいが，DEMIO の白色やシルバー色は落札価格を上げるポジティブ要因であり，黒色は落札価格を下げるネガティブ要因である．なお緑色の評価に差が見られるが，個体数が少ないことが理由である．決定木の分割基準（式 (A.6)）において，該当するサンプル数が少ないほどゲイン G_m が変化しにくいため分割基準に選ばれにくく，XGBoost では過小評価されやすい．実際に緑色の DEMIO は他色に比べて流通台数は少ないが，旧型では緑色を基調色としていたため，偏回帰係数 $\hat{a}_p$ はプラスに寄与したと考えられる．なお車体色はダミー変数として扱うため，重回帰式では基準（シルバー色）からの相対的な影響力を示す．したがって基準

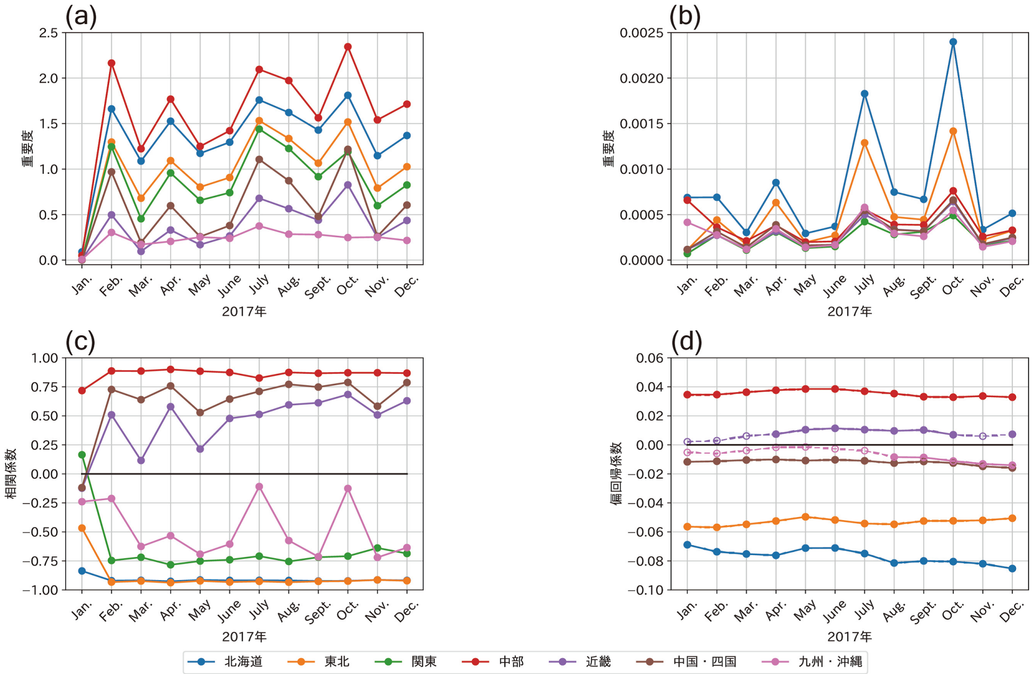

図 9　図 6 と同様，ただしオートオークション開催地域の要因分析（PRIUS の例）

Fig. 9　Same as Fig. 6, but factor analysis of auto-auction areas in PRIUS.

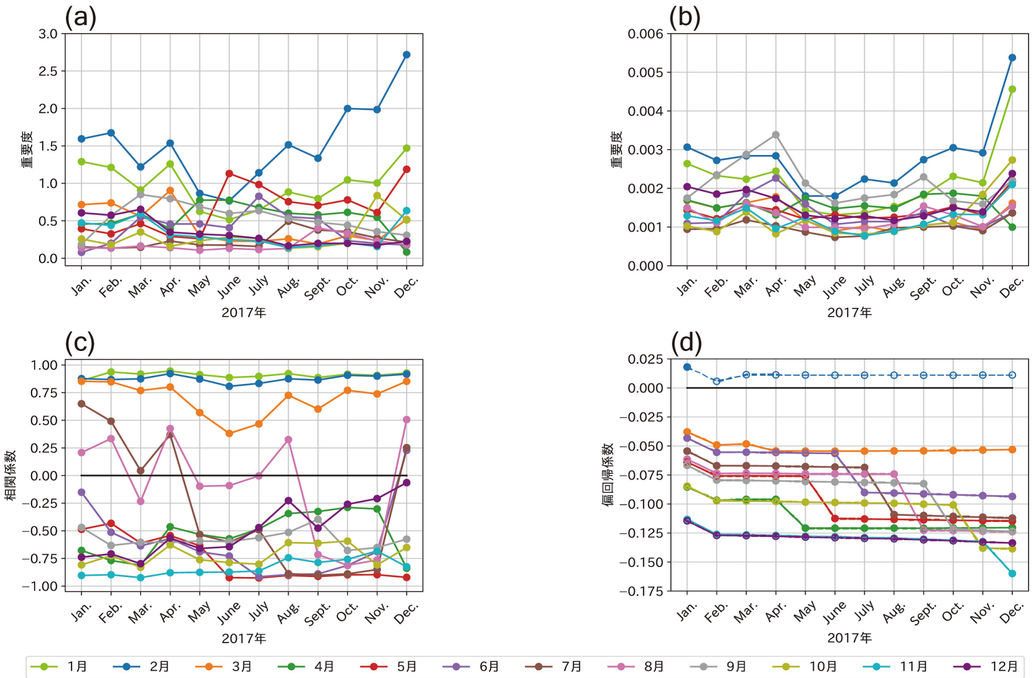

図 10　図 6 と同様，ただし開催月の要因分析（WagonR の例）

Fig. 10　Same as Fig. 6, but factor analysis of auto-auction months in WagonR.

を設けない XGBoost とは評価結果が異なる．

図 9 はオートオークション開催地域による影響力を示す．r_p や $\hat{a}_p$ より，中部地域は高値で落札される傾向にあり，北部の地域ほど安値で落札されやすい．なお $\hat{a}_p$ は基準（関東地域）からの相対評価であり，中部地域は 4%程度の増加要因，北海道地域は 8%程度の減価要因である．このように落札価格には地域性が存在する．

図 10 はオートオークション開催月による影響力を示す．r_p より，年度末の 1〜3 月は高値で落札される傾向にあり，新年度を迎えた 4〜6 月は安値で落札されやすい．これは中古車の新学期需要により，年度末に買い手が増えるためだと考えられる．さらにボーナス支給月（6 月と 12 月）前の需要減少により，9〜12 月も安値で落札される傾向にある．なお $\hat{a}_p$ は基準（1 月）からの相対評価であるが，おおむね同様の傾向を示している．このように落札価格は地域性に加え，季節性も存在する．

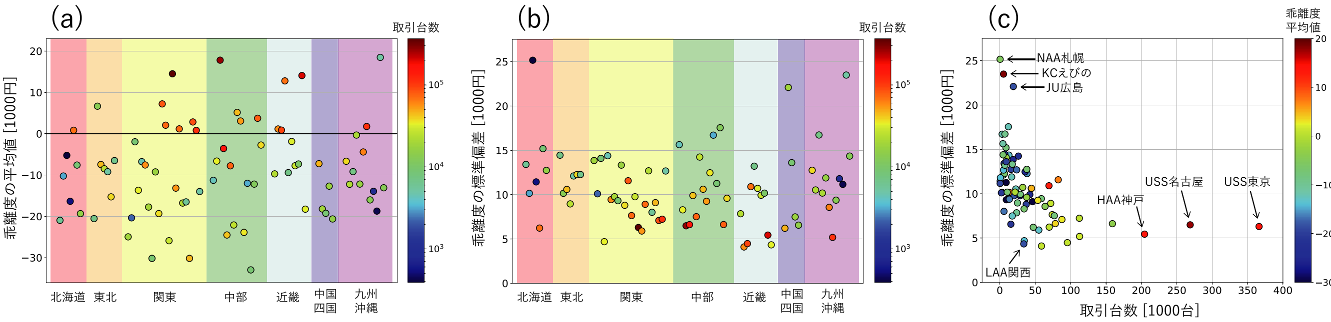

図 11　表 5 の平均値と標準偏差の関係（全 78 会場）

Fig. 11　Mean value and standard deviation shown in Table 5, but in all auto-auction venues.

表 5　会場別の乖離度 $\phi_v(t)$［1,000 円］

Table 5　Deviance $\phi_v(t)$ [1,000 yen] in major auto-auction venues.

会場名	1 月	2 月	3 月	4 月	5 月	6 月	7 月	8 月	9 月	10 月	11 月	12 月	平均値	標準偏差
USS 札幌	−0.8	7.6	12.0	5.2	3.8	−3.8	5.7	−0.8	1.6	−1.8	−10.0	−8.8	0.83	6.23
USS 東北	−1.8	−3.5	5.8	−20.1	−29.4	−14.2	−7.8	−13.0	11.3	−5.7	−1.1	−10.0	−7.45	10.59
USS 東京	27.3	22.9	13.2	4.1	8.2	9.6	18.7	18.3	17.0	13.0	10.2	11.6	14.50	6.30
USS 名古屋	31.1	29.8	18.9	10.0	10.8	12.7	20.7	16.6	17.6	17.8	12.4	15.2	17.79	6.49
HAA 神戸	14.7	22.9	17.6	7.7	9.4	7.0	19.8	17.5	18.6	16.9	5.8	10.9	14.07	5.44
TAA 広島	−3.5	−4.4	−0.9	−18.5	−10.0	−12.2	−6.9	−15.2	−7.2	0.8	2.5	−11.4	−7.25	6.21
USS 九州	3.0	13.7	0.7	−4.7	−4.8	−2.1	2.6	8.8	4.1	−1.9	−1.1	2.6	1.74	5.18

7.　落札価格の割安・割高評価

前章までの落札価格の推定は，主に中古車の買取査定モデルとしての活用を想定している．さらに本章では，買取査定モデルの応用として，落札価格の割安・割高評価を試みる．

まず落札価格の推定には，5 章で有用性を検証した XGBoost を適用する．毎月において前月から過去 2 年間の落札実績に基づいて式 (3) により関数 F を学習し，式 (4) により当月の中古車 j の推定落札価格 $\hat{y}_j$ を得る．これを検証期間において毎月繰り返す．

割安・割高評価において，オートオークション会場ごとで分析する場合は，ある月 t に会場 v で落札された中古車 j の集合を $J_v(t)$ と表記すると，

$$\phi_v(t) = \frac{1}{|J_v(t)|} \sum_{j \in J_v(t)} (y_j - \hat{y}_j) \tag{8}$$

によって乖離度 $\phi_v(t)$ を算出する．分析対象は日本全国の 78 会場（$v = 1 \sim 78$）とする．

推定値 $\hat{y}_j$ を算出する際に，XGBoost には個車状態・取引地域・取引月の説明変数（表 2）を与えているが，個別会場に関する情報は与えていない．したがって査定モデルは会場の違いによらない汎用モデルであるが，モデル残差である乖離度 $\phi_v(t)$ には会場 v の特殊性が残る．そこで，$\phi_v(t) > 0$ であるほど割高に落札されやすい会場であり，$\phi_v(t) < 0$ であるほど割安に落札されやすい会場と見なす．なお前章で確認したように，落札価格は地域性や季節性を持つが，これらの影響は XGBoost の出力値 $\hat{y}_j$ で調整される．

さらに車種ごとでも評価する場合は，中古車 j の集合に対象車種 c を条件に加えて $J_{v,c}(t)$ と標記すると，

$$\phi_{v,c}(t) = \frac{1}{|J_{v,c}(t)|} \sum_{j \in J_{v,c}(t)} (y_j - \hat{y}_j) \tag{9}$$

によって乖離度 $\phi_{v,c}(t)$ を算出する．分析対象は表 1 に示す 20 車種（$c = 1 \sim 20$）とする．

表 5 および**図 11** で，会場別の乖離度 $\phi_v(t)$ を示す．表 5 では誌面の都合により各地域で取引台数が多い 7 会場に限定し，図 11 で全会場の結果を示す．なお統計的なサンプル数を増やすために検証期間を 2013 年～2017 年の 5 年間に拡大し，月 t ごとに $\phi_v(t)$ を統合して平均化した．図 11 より，取引台数が多い会場ほど乖離度 $\phi_v(t)$ の平均値は 0 に近く，標準偏差も小さく安定的である．取引台数は会場の規模と同義であるため，大規模会場ほどオークション参加者が多く，合理的な価格形成がされやすいと考えられる．一方，取引台数が少ない小規模会場では，乖離度 $\phi_v(t)$ の標準偏差が大きく，特に $\phi_v(t) < 0$（割安）になりやすい．この理由として，一般的なオートオークションは競り上げ方式であるため，参加者が少ないほど安値で成約しやすい[*1]．そこで小規模会場で割安に落札できる機会を窺い，大規模会場で転売するような自己売買（ディーリング）が考えられる．なお小規模会場においても割高な落札は起こりうるが，ディーリング目的においては入札しなければよい．

[*1] 逆に USS 東京や USS 名古屋など一部の巨大規模会場では，参加者が多すぎるため $\phi_v(t) > 0$（割高）になりうると推察される．

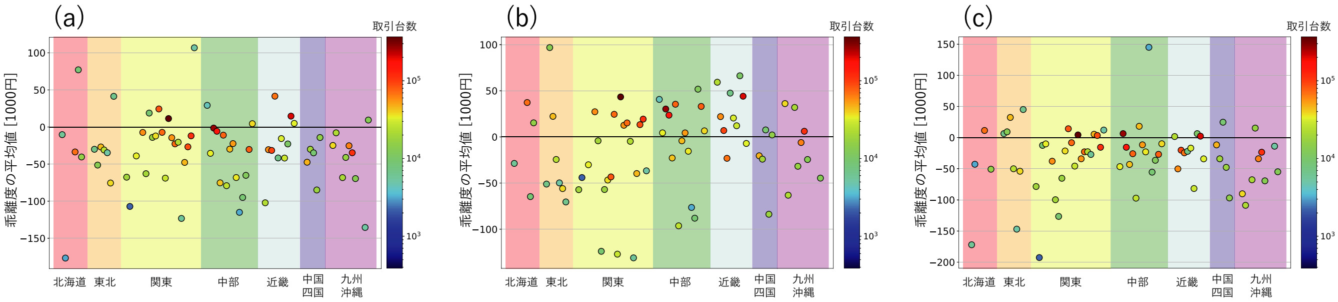

図 12　ESTIMA の乖離度 $\phi_{v,c}(t)$（全 78 会場）(a) 4 月，(b) 8 月，(c) 12 月

Fig. 12　Deviance $\phi_{v,c}(t)$ of ESTIMA in all auto-auction venues held in (a) April, (b) August, (c) December.

車種の違いを考慮した異常度 $\phi_{v,c}(t)$ として，**図 12** に ESTIMA の結果を示す．表 5 と同様に検証期間を 2013 年～2017 年とし，月 t ごとに $\phi_{v,c}(t)$ を統合して平均化した．会場の規模による傾向は図 11 (a) と同様であり，異なる季節においても大規模会場では $\phi_{v,c}(t) \simeq 0$ に近いが，小規模会場ほど $\phi_{v,c}(t) < 0$ になりやすい．

8.　まとめ

全国のオートオークションで落札された中古車ビッグデータに基づいて，重回帰式ベースのヘドニックアプローチや非線形モデルの XGBoost により個車価格推定モデルを構築した．中古車の各属性は落札価格に対して非線形的かつ間接的に影響することにより，XGBoost を適用することで推定精度を向上できた．要因分析においては，落札価格に寄与する各属性の重要度やポジネガ極性を評価した．重回帰式により線形的な直接効果を，XGBoost により非線形的な間接効果を含めて総合的に評価した．さらに機械学習の説明力を高める工夫として SHAP を導入し，オートオークション落札価格には季節性や地域性が存在することを確認した．最後に価格推定モデルの応用として，実際の落札価格と推定値の乖離を調べることで，小規模会場ほど割安で落札されやすい傾向を確認した．これにより本研究の価格推定モデルは，中古車の買取査定だけでなく，オートオークション会場を横断するディーリングにも活用できると示唆される．

謝辞　本研究の遂行にあたり有益なご助言をいただいた，下山力三氏，長谷川恵理子氏，平野大河氏，ならびに鈴木研究室修了生の櫻井大宙氏，速見勇磨氏，山下梨瑳氏に感謝申し上げます．本稿の内容はすべて筆者個人の見解であり，所属機関の公式見解ではありません．なお本研究の一部は JSPS 科研費（20K11969）の助成により行われました．

参考文献

[1]　Court, T.A.: Hedonic Price Indexes with Automotive Examples, *The Dynamics of Automobile Demand*, pp.98–119 (1939).

[2]　Griliches, Z.: Hedonic Price Indexes for Automobiles–An Econometric of Quality Change, *The Price Statistics of the Federal Goverment*, pp.173–196 (1961).

[3]　太田　誠：ヘドニック・アプローチの理論的基礎，方法及び日本の乗用車価格への応用，季刊理論経済学，Vol.29, No.1, pp.31–55 (1978).

[4]　Harrison, D. and Rubinfeld, L.D.: Hedonic Housing Prices and the Demand for Clean Air, *Journal of Environmental Economics and Management*, Vol.5, No.1, pp.81–102 (1978).

[5]　Noland, C.: Assessing Hedonic Indexes for Housing, *Journal of Financial and Quantitative Analysis*, Vol.14, No.4, pp.783–800 (1979).

[6]　国土交通省：不動産価格指数の作成方法，入手先 ⟨https://www.mlit.go.jp/common/001360416.pdf⟩（参照 2020-09-15）.

[7]　清水千弘：ビッグデータで見る不動産価格の決まり方，日本不動産学会誌，Vol.31, No.1, pp.45–51 (2017).

[8]　大和大祐，野村眞平：SUUMO でのビッグデータ活用事例，日本不動産学会誌，Vol.31, No.1, pp.78–83 (2017).

[9]　Meng, M.S., Liu, J.L., Kuritsyn, M., et al.: Price Determinants on Used Car Auction in Taiwan, *International Journal of Asian Social Science*, Vol.9, No.1, pp.48–58 (2019).

[10]　Shimizu, C., Nishimura, G.K. and Karato, K.: Nonlinearity of Housing Price Structure–The Secondhand Condominium Market in Tokyo Metropolitan Area, *International Journal of Housing Market and Analysis*, Vol.7, No.3, pp.459–488 (2014).

[11]　Ribeiro, M.T., Singh, S. and Guestrin, C.: "Why Should I Trust You?" Explaining the Predictions of Any Classifier, *Proc. KDD 2016*, pp.1135–1144 (2017).

[12]　Lundberg, S.M. and Lee, S.I.: A Unified Approach to Interpreting Model Predictions, *Proc. Advances in Neural Information Processing Systems* (*NIPIS*), pp.4768–4777 (2017).

[13]　Lu, J.: The Value of a South-facing Orientation–A Hedonic Pricing Analysis of the Shanghai Housing Market, *Habitat International*, Vol.81, pp.24–32 (2018).

[14]　Chen, T. and Guestrin, C.: XGBoost–A Scalable Tree Boosting System, *Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp.785–794 (2016).

[15]　xgboost developers: XGBoost Documentation, available from ⟨https://xgboost.readthedocs.io/en/latest/python/python_api.html⟩(accessed 2021-01-01).

付　　録

A.1　XGBoost [14] の概要

学習データにおける目的変数を y_i（$i=1,2,\cdots,n$）とし，説明変数を $\boldsymbol{x}_i=\{x_{1,i},x_{2,i},\cdots,x_{P,i}\}$ とする．これらを用いて，まず決定木 $f^{(1)}$ を 1 つ構築し，$\hat{y}_i^{(1)}=f^{(1)}(\boldsymbol{x}_i)$ のように推定値を得る．ここで $\epsilon_i^{(1)}=y_i-\hat{y}_i^{(1)}$ が推定誤差となる．

次に，この推定誤差 $\epsilon_i^{(1)}$ を目的変数とした決定木 $f^{(2)}$ を構築する．推定値は $\hat{\epsilon}_i^{(2)}=f^{(2)}(\boldsymbol{x}_i)$ となるため，

$$\hat{y}_i^{(2)}=\hat{y}_i^{(1)}+\hat{\epsilon}_i^{(2)} \tag{A.1}$$

のように推定値を補正する．これを K 回繰り返し，

$$\hat{y}_i^{(K)}=\hat{y}_i^{(1)}+\hat{\epsilon}_i^{(2)}+\cdots+\hat{\epsilon}_i^{(K)} \tag{A.2}$$

を最終的な推定値 $\hat{y}_i=F(\boldsymbol{x}_i)$ とする．ここで関数 F が XGBoost に相当する．

XGBoost を構成する際に，2 つ目以降の決定木（$k=2\sim K$）では，以下の目的関数 $L^{(k)}$ を最小化するように学習させる．

$$\begin{aligned}L^{(k)}&=\sum_{i=1}^{n}\left[g_i f^{(k)}(\boldsymbol{x}_i)+\frac{1}{2}h_i[f^{(k)}(\boldsymbol{x}_i)]^2\right]\\&\quad+\gamma M+\frac{1}{2}\lambda||\omega||^2\end{aligned} \tag{A.3}$$

ここで $g_i=\partial_{\hat{y}_i^{(k-1)}}l(y_i,\hat{y}_i^{(k-1)})$，$h_i=\partial^2_{\hat{y}_i^{(k-1)}}l(y_i,\hat{y}_i^{(k-1)})$ であり，それぞれ損失関数 l の 1 次，2 次勾配を表す．なお，M は木の終端ノードの数，γ と λ は罰則パラメータ，ω は決定木 $f^{(k)}$ が返す補正量である．

学習データ $\boldsymbol{x}_i$ が m 番目の最終ノードに到達したとき，決定木は $f^{(k)}(\boldsymbol{x}_i)=\omega_m$ のように返すとすると，式 (A.3) を最小化する ω_m の最適解は，以下のように導出される．

$$\omega_m=-\frac{\sum_{i\in I_m}g_i}{\sum_{i\in I_m}h_i+\lambda} \tag{A.4}$$

ここで I_m は，m 番目の最終ノードに到達した学習データ i の集合である．

式 (A.4) を式 (A.3) に代入すると，$||\omega||^2=\sum_{m=1}^{M}\omega_m^2$ より，

$$L^{(k)}=-\frac{1}{2}\sum_{m=1}^{M}\frac{\left(\sum_{i\in I_m}g_i\right)^2}{\sum_{i\in I_m}h_i+\lambda}+\gamma M \tag{A.5}$$

となる．親ノード m を子ノード L と R に分岐する際，分割後の目的関数を $L_2^{(k)}$ と書くと，分割によって得られる $L^{(k)}$ の減少量（ゲイン）G_m は次式となる．

$$\begin{aligned}G_m&=L^{(k)}-L_2^{(k)}\\&=\frac{1}{2}\left[\frac{\left(\sum_{i\in I_L}g_i\right)^2}{\sum_{i\in I_L}h_i+\lambda}+\frac{\left(\sum_{i\in I_R}g_i\right)^2}{\sum_{i\in I_R}h_i+\lambda}-\frac{\left(\sum_{i\in I_m}g_i\right)^2}{\sum_{i\in I_m}h_i+\lambda}\right]\end{aligned} \tag{A.6}$$

ここで I_L と I_R は分割後の子ノードに含まれるデータの集合を表し，$I_m=I_L\cup I_R$ である．このゲイン G_m が最大となるように親ノード m の分割条件を網羅的に探索して決定する．このノード分割を終了条件に至るまで再帰的に繰り返すことで，決定木 $f^{(k)}$ を成長させる．

A.2　SHAP [12] の概要

SHAP（SHapley Additive exPlanations）は機械学習モデル F の出力値 $\hat{y}_i$ に対する根拠を説明する手法の 1 つであり，その説明モデル g を以下のように設定する．

$$g(\boldsymbol{x}_i)=\phi_0+\sum_{p=1}^{P}\phi_p(\boldsymbol{x}_i,F)\cdot h(x_{p,i}) \tag{A.7}$$

ここで関数 $h(\cdot)$ は入力値が存在したら 1 を，存在しなければ 0 を出力するものとする．なお ϕ_0 は平均的な出力値（$\frac{1}{n}\sum_{i=n}^{n}\hat{y}_i$）とし，$\phi_p$ の加算によって出力値 $\hat{y}_i=F(\boldsymbol{x}_i)$ への貢献度を表したい．そこで $F(\boldsymbol{x}_i)=g(\boldsymbol{x}_i)$ になるように SHAP 値 $\phi_p(\boldsymbol{x}_i,F)$（$p=1,2,\cdots,P$）を導くと[*2]，

$$\begin{aligned}\phi_p(\boldsymbol{x}_i,F)\;=\;&\sum_{\omega_p\subseteq\Omega_p}\frac{|\omega_p|!\,(|\Omega_p|-|\omega_p|-1)!}{|\Omega_p|!}\\&\cdot(F(\{x_{\tilde{p},i}|\tilde{p}\in\omega_p\})-F(\{x_{\tilde{p},i}|\tilde{p}\in\omega_p\backslash p\}))\end{aligned} \tag{A.8}$$

が得られ，ゲーム理論における Shapley 値と一致する．なお Ω_p は説明変数 p の全体集合であり，ω_p はその部分集合である．

工藤 大輝

令和元年茨城大学工学部知能システム工学科卒業．同年 4 月より同大学大学院理工学研究科機械システム工学専攻博士前期課程に進学．学生フォーミュラ日本大会 2019 においてベスト・サスペンション賞を受賞する等自動車に関する機械工学や，統計分析や機械学習によるデータサイエンスに従事．

福西 亮介

株式会社 KINTO 所属．自動車に関する事業やプロジェクトに従事．

*2　導出条件の詳細については文献 [12] を参照されたし．

黛 広樹

株式会社プロトコーポレーション所属．IT 部門執行役員．グーネット中古車（Goo-net）等 IT を活用した事業やプロジェクトに従事．

鈴木 智也

平成 17 年東京理科大学大学院理学研究科物理学専攻博士課程修了．理学博士．同年東京電機大学工学部電子工学科助手，平成 18 年同志社大学工学部情報システムデザイン学科専任講師，平成 21 年茨城大学工学部知能システム工学科准教授，平成 28 年より同大学教授．主に，金融業務に関する機械学習やデータサイエンスに従事．