経営課題にAIを!ビジネスインフォマティクス研究会 (第10回)
JSAI Special Interest Group on Business Informatics (SIG-BI #10)

機械学習によるオートオークション落札価格の予測

Price Prediction of Used Cars at Auto Auction by Machine Learning Methods

櫻井 大宙 [1*]　速見 勇磨 [1]　山下 梨瑳 [1]　長谷川 恵理子 [2]
下山 力三 [2]　福西 亮介 [2]　黛 広樹 [2]　鈴木 智也 [1, 3]
Hiromichi Sakurai[1], Yuma Hayami[1], Risa Yamashita[1], Eriko Hasegawa[2],
Rikizou Shimoyama[2], Ryosuke Fukunishi[2], Hiroki Mayuzumi[2] and Tomoya Suzuki[1, 3]

[1] 茨城大学大学院理工学研究科
[1] Graduate School of Science and Engineering, Ibaraki University
[2] 株式会社プロトコーポレーション
[2] PROTO CORPORATION
[3] コラボウィズ株式会社, [3] CollabWiz, Inc.

Abstract: Buyers of used cars have to predict their dealing prices at auto auction held after about one month, but it is very difficult to predict them because each condition of used cars is completely different such as mileage, model year, body color, etc. For this reason, we propose two prediction methods: as the first one, we consider the median of dealing prices in each car model as a base price and predict its future price by time-series models: ARIMA and SARIMA. After that, the predicted bace price is converted into the individual price of each used car by the machine learning method that learned the relationship between the condition of used cars including individual prices and the bace price. As the second method, we adopt the deep learning approach to directly predict individual future prices of used cars without using the base price, but using all the information attached to each used car as explanatory variables. To verify the usefulness of our proposed methods for a used car assessment system, we performed some prediction tests using the real auto-auction price data.

1　はじめに

中古車はユーザーから買取店へ流れた後，主にオートオークション [1] を通じて販売店に売却される．買取店の立場に立てば，買取からオートオークションに出品するまで一ヶ月は要するため，翌月の競買価格を予想して買取査定価格を設定する必要がある．そこで株式会社プロトコーポレーション (以下，プロト社) では，翌月に開催されるオートオークションでの落札価格を考慮した査定システムを製作し，中古車買取店が査定価格を決定する際の情報提供を行っている．

しかし中古車の流通価格を予測することは大変難しい．その理由として，中古車は走行距離や車検残月といった特徴が個車ごとに異なり，唯一無二の存在である．そこでプロト社の査定システム [2] では，まず対象個車と同一車種の落札実績を踏まえ，翌月の該当車種の指定条件の価格（代表価格）を推論し，さらに対象個車の状態（説明変数）に基づいて個車の落札価格へ変換するといった，2 段階のプロセスを経る．各プロセスでは主にパラメータを与えて機械的に推論するが，不合理な結果を訂正するために人間が有する柔軟な推論能力を駆使し，フレーム問題に対処する場合もある．しかし本研究では，全て機械的に推論することで全自動かつ主観が介在しない予測システムを構築したい．

まずはプロト社のように 2 段階のプロセスに分割し，代表価格の予測には時系列モデルを，個車価格への変換には機械学習モデルを適用する．しかし代表価格の予測誤差がその後の変換に悪影響を及ぼすため，個車の落札価格を直接予測する深層学習モデルについても検討する．直接予測では個車の有する全特徴量を説明変数とするため高次元の機械学習となるが，積層オートエンコーダを事前学習に適用することで次元圧縮（特徴選択）を行う．

*連絡先： 茨城大学大学院理工学研究科 茨城県日立市中成沢町 4-12-1 E-mail: 17nm920a@vc.ibaraki.ac.jp

2 問題設定と用いるデータ

中古車取引の流れは次の 4 つに分類される．(1) 買取店が顧客から中古車を買い取る，(2) 買い取った中古車を買取店がオートオークションに出品する，(3) 販売店が中古車を落札する，(4) 販売店が顧客に中古車を売却する．本研究では中古車買取店の立場に立ち，(3) での落札価格を考慮しつつ，(1) にて合理的な買取価格を決めるための方法論を検討する．もし買取価格を高く見積もれば買取店の利益が減り，低く見積もれば顧客が他店に流れる可能性がある．また買い取りからオートオークションへの出品まで整備等に一週間から一ヶ月程度のタイムラグを要すため，(1) の時点で (3) での落札価格を予測する必要がある．

本研究の予測対象として,「セダン」「コンパクト」「ミニバン」「ＳＵＶ」「軽自動車」の 5 カテゴリから,「ＬＳ」「プリウス」「フィット」「デミオ」「セレナ」「パジェロ」「フォレスター」「タント」「ワゴンＲ」の 9 車種を選出した．プロト社が有するオートオークションデータは，オートオークション会場が検査した個車の状態を表す多変量データで構成されている．その一例として「落札価格」「走行距離」「車検満了年月」「修復歴」「内装評価」などが挙げられる．またマスターデータとして,「車種」「モデル」「型式」「グレード」に応じた「キーコード」「公称性能」「オプション装備」「新車価格」に関する情報も利用できる．このキーコードが自動車を分類する最小単位であるため，オートオークションデータからキーコードを紐付けてマスターデータから情報を抽出する．

3 代表価格の時系列予測

3.1 時系列予測モデルの適用

オートオークションにて中古車が落札されるのは買取店が査定買取してから一ヶ月程度先であることが多く，買取店は中古車を査定するときに翌月の落札価格を予測する必要がある．この予測には査定当月までのオークション落札情報を参考にするが，前述のとおり対象車両と同一条件の過去実績は存在しないため，統計的に推論することができない．そこで本章では充分なサンプル数を確保するために，個車ではなく車種毎に落札価格をトラッキングする．また車種の代表価格として，落札価格の中央値を機械的に用いる．

次に，翌月の代表価格（中央値）を予測するために，過去の代表価格を用いて時系列モデルを学習する．中古車は流行などの「トレンド要因」や新学期需要などの「季節性要因」が影響すると考えられる．これらの非定常要因に対する時系列モデルとして，本研究では自己回帰和分移動平均 (ARIMA) モデルと季節自己回帰和分移動平均 (SARIMA) モデルを適用する．

時刻 t におけるある車種の代表価格（中央値）を X_t と書くと，ARMA(p,q) モデルは次式となる．

$$X_t = \sum_{i=1}^{p} a_i X_{t-i} + \epsilon_t + \sum_{i=1}^{q} b_i \epsilon_{t-i} \tag{1}$$

ここで，a_i は過去の実現値 X_{t-i} における時間記憶力，b_i は過去のホワイトノイズ ϵ_{t-i} における時間記憶力を示す．ARMA モデルは定常性を仮定するため，上記の非定常要因を表現できない．X_t のトレンド要因について d 階差分をとり，季節性要因について D 階季節差分を取ることで定常化処理を施す．なお X_t は月次データであるため，季節周期は $s = 12$（12ヶ月）とする．その後，定常化済み $\Delta^d \Delta_s^D X_t$ について式 1 の ARMA モデルを適用し，推定された $\Delta^d \Delta_s^D \hat{X}_{t+1}$ を逆算することで $\hat{X}_{t+1}$ を得る．式 1 では時間系列に関する記憶性を表現できるため，ARIMA(p,d,q) モデルと等価である．

しかし季節周期 s をまたぐ時間記憶性も表現するには，次式の SARIMA$(p,d,q)(P,D,Q)[s]$ モデル [3] を用いる．

$$\begin{aligned}\Delta^d \Delta_s^D X_t = &\sum_{i=1}^{p} a_i \Delta^d \Delta_s^D X_{t-i} + \sum_{I=1}^{P} A_I \Delta^d \Delta_s^D X_{t-sI} \\ &- \sum_{i=1}^{p}\sum_{I=1}^{P} a_i A_I \Delta^d \Delta_s^D X_{t-sI-i} + \epsilon_t + \sum_{j=1}^{q} b_j \epsilon_{t-j} \\ &+ \sum_{J=1}^{Q} B_J \epsilon_{t-sJ} + \sum_{j=1}^{q}\sum_{J=1}^{Q} b_j B_J \epsilon_{t-sJ-j}\end{aligned} \tag{2}$$

上記と同様に，推定された $\Delta^d \Delta_s^D \hat{X}_{t+1}$ を逆算することで $\hat{X}_{t+1}$ を得る．ただしモデル構造が複雑であるため，学習に伴うデータ不足および過学習の問題が懸念される．モデル次数については AIC を最小化するように決定するが，過学習の観点からよりシンプルな ARIMA モデルの方が機能する可能性もある．

3.2 予測実験

2011 年 10 月から予測対象の前月までの代表価格を用いて時系列モデルをオンライン学習し，2016 年 1 月から 2017 年 9 月を予測対象とした．予測精度の評価には，正規化平均二乗誤差 (NMSE) と正答月数を用いた．NMSE は次式で得られる．

$$\mathrm{NMSE}(\hat{X}) = \frac{\sqrt{\sum_{t=1}^{T}(X_t - \hat{X}_t)^2}}{\sqrt{\sum_{t=1}^{T}(X_t - \bar{X})^2}} \tag{3}$$

表 1: ARIMA モデルによる代表価格の予測結果

カテゴリ	車種	NMSE	正答月数
セダン	ＬＳ	0.559	17
セダン	プリウス	0.621	19
コンパクト	フィット	0.552	15
コンパクト	デミオ	0.546	14
ミニバン	セレナ	0.629	18
ＳＵＶ	パジェロ	1.613	14
ＳＵＶ	フォレスター	1.062	9
軽自動車	タント	0.454	19
軽自動車	ワゴンＲ	0.913	15

表 2: SARIMA モデルによる代表価格の予測結果

カテゴリ	車種	NMSE	正答月数
セダン	ＬＳ	0.466	20
セダン	プリウス	0.797	18
コンパクト	フィット	0.631	15
コンパクト	デミオ	0.515	15
ミニバン	セレナ	0.666	18
ＳＵＶ	パジェロ	1.127	17
ＳＵＶ	フォレスター	1.004	11
軽自動車	タント	0.451	19
軽自動車	ワゴンＲ	0.850	17

ここで，T は予測対象の期間の長さ，X_t は時刻 t における代表価格の真値，$\hat{X}_t$ は時刻 t における代表価格の予測値，$\bar{X}$ は期間 $t = 1, 2, \cdots, T$ における X_t の平均値を示す．正答月数は，$\hat{X}_t$ が X_t の $\pm 10\%$ に収まった場合を正解とみなし，予測対象の 21ヶ月間 $(T = 21)$ のうち正解した月数を示す．

予測結果を表 1,2 および図 1~3 に示す．セダンは ARIMA モデルと SARIMA モデルの両方において正答月数が多く，NMSE も低い値となった．理由として，セダンのトレンドは比較的落ち着いており，季節性についても変動幅が他車種より小さい．一方，趣味性が高いＳＵＶはトレンド変化が激しく，季節性変動幅も大きいため予測が難しい．その結果，正答月数は少なく，NMSE も高い傾向にある．特に図 2 において，予測値が真値を 1ヶ月遅れて追いかけている期間が見られる．なお ARIMA モデルと SARIMA モデルの比較においては，特記するほどの差は見受けられない．

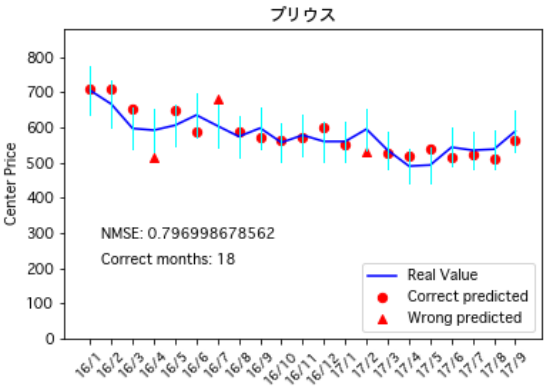

図 1: SARIMA によるプリウスの代表価格の予測推移

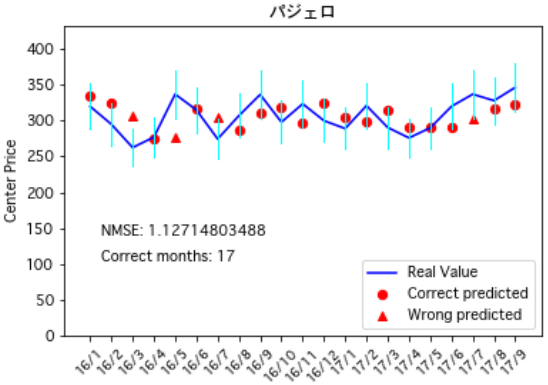

図 2: SARIMA によるパジェロの代表価格の予測推移

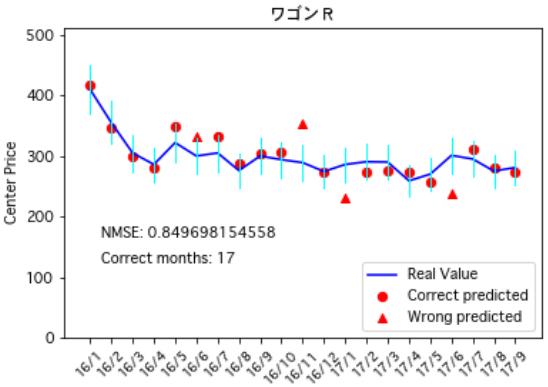

図 3: SARIMA によるワゴンＲの代表価格の予測推移

4　機械学習による個車価格への変換

4.1　ランダムフォレストの適用

次に，代表価格 (中央値) の予測値を個車価格へ変換する方法を検討する．個車の特徴量と代表価格の予測値を説明変数とし，個車の落札価格を被説明変数にすれば，両変数の関係を過去の実績に基づいて機械学習すれば良い．その機械学習モデルとして，本章ではランダムフォレスト [4] を用いる．

その学習手順として，まず N 個の学習データ $(\boldsymbol{x}_i, y_i)$, $(i = 1, 2, \cdots, N)$ から復元抽出し，同サイズ N のデータ集合 $W^{(b)}$ を B 組生成する．次に，それぞれのデータ集合 $W^{(b)}$, $(b = 1, 2, \cdots, B)$ を用いて回帰木 $RT^{(b)}$ を学習する． 回帰木は単体でも連続的な被説明変数の予測に適用できるが，ランダムフォレストは全ての回帰木の予測値を平均化することで予測性能を高めるアンサンブル学習 [5] の一種である．各回帰木の独立性が高いほど予測性能が高まることは集合知の観点（多様性予測定理）から説明できる [6].

表 3: ランダムフォレストによる個車落札価格の推定結果

カテゴリ	車種	正答率	決定係数 R^2
セダン	ＬＳ	0.3068	0.8802
セダン	プリウス	0.4338	0.8921
コンパクト	フィット	0.2332	0.8933
コンパクト	デミオ	0.3222	0.9433
ミニバン	セレナ	0.3186	0.9328
ＳＵＶ	パジェロ	0.3291	0.9122
ＳＵＶ	フォレスター	0.4076	0.9549
軽自動車	タント	0.4724	0.9516
軽自動車	ワゴンＲ	0.4269	0.9454

4.2　変換実験

学習データの期間は 2016 年 1 月から 2016 年 12 月まで，テストデータの期間は 2017 年 1 月から 2017 年 9 月までとする．まず学習データを用いてランダムフォレストを構築する．その際には，個車情報として「車体色」「年落ち度」「走行距離」「車検残月」「評価点」「新車価格」「ナビの有無」「レザーシートの有無」「サンルーフの有無」を用い，さらに当該車種の代表価格（中央値）を説明変数とする．被説明変数には同車両の「落札価格」とすることで，車種毎に代表価格から個車の落札価格に変換するルールを学習する．その後，代表価格を 3 章で得た予測値に変更し，これを学習済みランダムフォレストに入力することで，個車の落札価格の推定値を得る．この推定はテストデータに対して実施する．

推定精度を評価する指標として，正答率と決定係数を用いる．正答率は，推定値 $\hat{y}_n$ が真値 y_n の $\pm 10\%$ 以内に収まった場合を正解とし，その回数を推定実施回数 N で割ったものである．さらに決定係数 R^2 は次式で得られる．

$$R^2 = 1 - \frac{\sum_{n=1}^{N}(y_n - \hat{y}_n)^2}{\sum_{n=1}^{N}(y_n - \bar{y})^2} \tag{4}$$

ここで $\bar{y}$ は真値 y_n の平均値である．決定係数 R^2 は 1 に近いほど回帰問題として当てはまりが良いことを意味する．

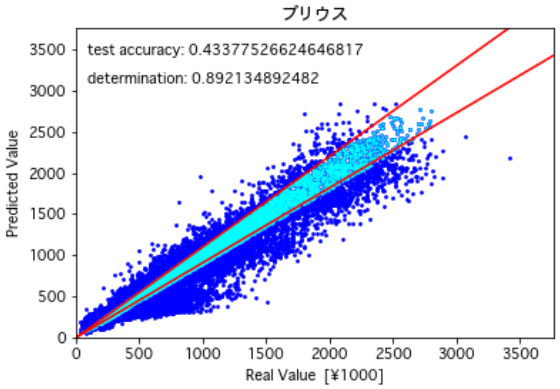

図 4: 個車落札価格の真値と推定値の相関図 (プリウスの場合)

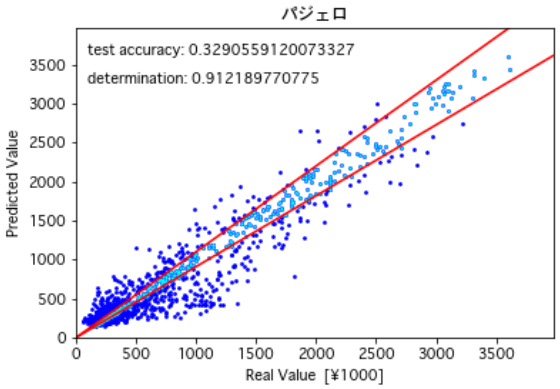

図 5: 個車落札価格の真値と推定値の相関図 (パジェロの場合)

推定結果を表 3 および図 4,5 に示す．いずれの車種も決定係数 R^2 は 0.9 に近い値を示しており，回帰問題としての推定精度は高い．しかし正答率は 0.30 前後と低く，真値の ±10% 以内を正解とみなす基準は非常に厳しい．

また図 4 と図 5 に示すように，低価格帯において誤差が大きい傾向にある．この理由として，グレードや年式などで分類せずに学習したため，高価格帯と低価格帯の学習誤差が同一に扱われたためと考えられる．真値と誤差の比率を考えれば，低価格帯はこの比率が大きくなりやすい．これが正答率を下げる要因になるため，今後は絶対誤差でなく「誤差率」を最小化するようにモデルパラメータを機械学習する方法を検討する．また，グレードや年式などで学習データを分割する方法も有効かもしれないが，学習データ数の減少により機械学習自体の性能が低下する可能性もある．

5 深層学習による個車落札価格の直接予測

5.1 深層学習の適用

3 章と 4 章による予測プロセスにおいて，誤差が発生する要因は 3 点ある．まず車種毎の代表価格として中央値が妥当なのか疑問が残る．しかし機械的に代表価格を求めるルールを定めても，そのルールが完全に正しいとは限らない．次に代表価格に対する時系列予測についても予測誤差が生じ，さらに個車落札価格への変換においても変換誤差が生じる．そこで本章では，これらの誤差が介在しないよう代表価格を考慮せず，直接的に個車の落札価格を予測する．この予測に深層学習モデルを適用する．

深層学習モデルの構成として中間層を 3 層とし，各ニューロン数を下位層から順に $j = 40$, $k = 8$, $l = 4$ とする．なお過学習を抑えるべく，中間層第 1 層には消去率 40% のドロップアウトを適用する．活性化関数については，中間層第 1 層と第 3 層に正規化線形関数 (以下，ReLU) を用い，中間層第 2 層に恒等関数を用いる．なお本研究は回帰問題 (連続値の予測) であるため，出力層の活性化関数も恒等関数とする．結合強度の初期値については，ReLU を用いる層へ結合する場合は Xavier Glorot の初期値 [7] を，恒等関数を用いる層へ結合する場合は He の初期値 [8] を用いる．

中間層を多層にするほど勾配消失問題が起こるため，事前学習として積層オートエンコーダ [9] を適用する．各オートエンコーダにおいて情報圧縮することで，説明変数の特徴抽出を行う．その後ファインチューニングとして，深層学習モデル全体を誤差逆伝播法によって再学習する．その際には，教師信号に対する二乗和誤差を誤差関数 E とし，学習データをランダムに 1% 選ぶバッチ学習を 10000 回繰り返す．勾配降下法の最適化アルゴリズムには Adam[10] を用い，各パラメータを $\alpha = 0.001$，$\beta_1 = 0.9$，$\beta_2 = 0.999$ とした．ファインチューニングにおいては勾配消失問題により入力層付近の重みは更新されにくいが，事前学習において優先的に学習されているため重大な問題とはならない．

5.2 直接予測実験

学習データとテストデータの期間は 4.2 節と同様である．説明変数として「車体色」「年落ち度」「走行距離」「車検残月」「評価点」「新車価格」「ナビの有無」「レザーシートの有無」「サンルーフの有無」を用い,「個車落札価格」を被説明変数とする．4.2 説との違いは，代表価格を説明変数に含めない点と，被説明変数の値を「査定時点に戻して」学習する点にある．例えば，年落ち度は予測対象個車の年式とオートオークション開催月の差を意味する．そこで年落ち度を 1ヶ月引いて査定時点の情報を深層学習に入力することで，1ヶ月先のオークション開催時点の落札価格 (予測値) を出力する．

さらに各変数のスケール調整 (前処理) として,「車体色」は車種毎の出現率として定量化し，色の価値 (人気度) を表す.「年落ち度」や「走行距離」は車種毎に学習データの全車両の間で標準化する.「車検残月」は通常満期が 2ヶ年なので査定時点における車検残月を 24 で除しておく．なお新車の初車検は 3ヶ年なので値は 0 ∼ 1.5 をとりうる.「評価点」はオートオークション会場による価値評価であり，新古車は 2000 点，中古車は 300 点 ∼900 点で採点される．そのため元の評価点を 1000 で除しておく.「新車価格」は車種毎に標準化する．さらに「ナビの有無」「レザーシートの有無」「サンルーフの有無」については新車時の標準装備状態と同一であれば 0，不足していれば-1，増設されていれば+1 とする．

予測結果を表 4 および図 6,7 に示す．セダンについては，表 3 のランダムフォレストより高い予測精度を得ている．これは 3 章の時系列予測と同様に，トレンドや季節性変動が比較的小さいほど予測しやすく，さらに直接予測により誤差が介在する要因を減らした効果だと考えられる．決定係数 R^2 においても，9 車種中 5 車種において表 3 を上回る結果を得たが，正答率では 9 車種中 2 車種しか表 3 から向上できていない．この理由として，深層学習の説明変数にトレンドや季節性に関する情報を用いなかったため，これらの要因が強い車種においては前章の方法論の方が機能したと考えられる．そこで今後は，トレンドについては RNN や LSTM など時間構造に対応した機械学習モデルを導入し，季節性については月や曜日などをダミー変数で与える方法を検討する．

表 4: 深層学習による個車落札価格の予測結果

カテゴリ	車種	正答率	決定係数 R^2
セダン	ＬＳ	0.4205	0.9420
セダン	プリウス	0.4492	0.9306
コンパクト	フィット	0.2283	0.8915
コンパクト	デミオ	0.2762	0.9328
ミニバン	セレナ	0.3182	0.9422
ＳＵＶ	パジェロ	0.3093	0.9340
ＳＵＶ	フォレスター	0.3593	0.9707
軽自動車	タント	0.4086	0.9374
軽自動車	ワゴンＲ	0.3941	0.9400

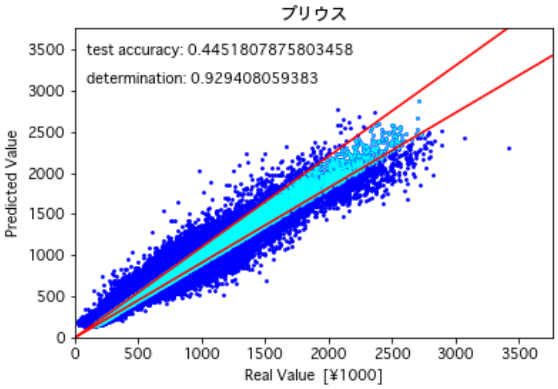

図 6: 個車落札価格の真値と予測値の相関図 (プリウスの場合)

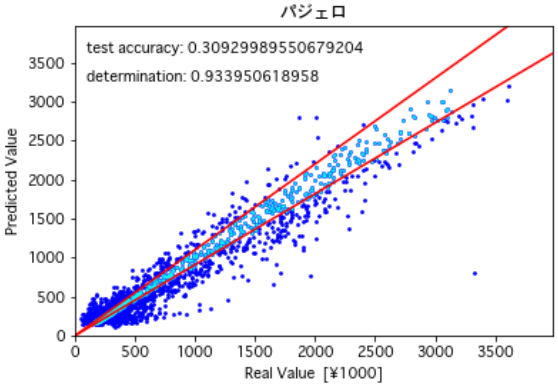

図 7: 個車落札価格の真値と予測値の相関図 (パジェロの場合)

6 まとめ

本研究では中古車買取店の立場に立ち，1ヶ月後の中古車の落札価格を予測した．その際に車種毎の代表価格 (中央値) を媒介する方法と，媒介せずに直接予測する方法を検討した．その結果，トレンドや季節性変動が大きい車種には前者が良く，小さい車種には後者が良かった．この理由として，後者の直接予測 (深層学習) は誤差の発生プロセスを最小限にできるが，前者 (時系列モデル) のように時間構造を考慮していない．そこで今後の課題として，RNN や LSTM でトレンドを考慮し，ダミー変数で季節性を表現し，さらに誤差率を機械学習の評価基準にすることで低価格帯における予測誤差の改善を目指す．

謝辞

本研究は株式会社プロトコーポレーションとの共同研究である．なお本研究の一部は文部科学省科学研究費基盤研究 (C)(No.16K00320) の助成により行った．

参考文献

[1] オートオークションのご案内 — Used car System Solutions: http://www.ussnet.co.jp/auction/ (参照日: 2018/8/1)

[2] DataLine 査定サービス: https://www.proto-g.co.jp/product/corporation/dataline_satei.html (参照日: 2018/9/1)

[3] 田中勝人: 現代時系列分析, 岩波書店, 2006.

[4] Trevor Hasite, et al.: 統計的学習の基礎 ―データマイニング・推論・予測―, 共立出版, 2014.

[5] Zhi-Hua Zhou: Ensemble Methods: Foundations and Algorithms, Chapman and Hall/CRC, 2012.

[6] 西垣通, 集合知とは何か, 中公新書, 2013.

[7] Xavier Glorot and Yoshua Bengio: “Understanding the difficulty of training deep feedforward neural networks,” Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, pp.249–256, 2010.

[8] Kaiming He, et al.: “Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification,” Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), pp.1026–1034, 2015.

[9] Yoshua Bengio, Pascal Lamblin, Dan Popovici, and Hugo Larochelle: “Greedy layer-wise training of deep networks,” Advances in Neural Information Processing Systems, vol.19, p.153, MITPress, 2007.

[10] Diederik P. Kingma and Jimmy Ba: “Adam: A Method for Stochastic Optimization,” arXiv preprint arXiv:1412.6980, 2015.