The 33rd Annual Conference of the Japanese Society for Artificial Intelligence, 2019

戸建住宅価格における機械学習を用いた2段階推計モデル

Two-Step estimation model using machine learning at detached house price

高橋 佑典 *1
Yusuke Takahashi

*1富士通クラウドテクノロジーズ株式会社
FUJITSU CLOUD TECHNOLOGIES LIMITED

This paper reports the result of trying to estimate the detached house price with a two step model. By handling the advantages of linear model and nonlinear model in combination, we can expect explanation possibility and estimation accuracy for model when estimating house price. Experiments compare the methods by machine learning and the effects of explanatory variables to be input to the nonlinear model.

1. はじめに

住宅価格の分析を行った従来の研究では，多くが線形モデルとして推定している．その背景として，シンプルなモデルであればモデルの推計価格に対して明確な説明が可能であるという点が大きい．しかし，専有面積や築年数などでは，住宅価格に与える影響が区間によって異なり，非線形になると考えられ，従来の研究でもその存在が示されている．また，線形モデルとして推定する場合，住宅価格を決定する上で主要な要因である築年数と建築年代といった多重共線性を持つ説明変数を同時に投入してしまうと，モデルの推計値が不安定になってしまうことも挙げられる．

そこで，先行研究のうち [Shimizu 14] では，非線形モデルとして推計するため，ノンパラメトリックなモデルである連続量 dummy モデル（DmM）と，AIC を評価指標とした Switching RegressionModel（SWR），一般加法モデル（GAM）が用いられ，非線形性を考慮した推計が行われている．非線形モデルを扱う課題として，モデルが複雑になり再現性が低下することが考えられる．

本研究では，線形モデルと非線形モデルの課題に対処するため，モデルを 2 段階に分けて戸建住宅における取引価格の推計を行う．1 段階目のモデルでは線形モデルとして推計し，モデルに対する説明性を担保する．2 段階目のモデルでは 1 段階目の誤差項を目的変数として，非線形なモデルによって推計を行う．本研究の最終的な目的は，線形モデルと非線形モデルを組み合わせた二段階のモデルによって推計することで，住宅価格の推計時に，説明性と推計精度を向上させることである．

実験では，最初に線形モデルと機械学習による非線形モデルを作成し，誤差率の分布を比較することで，非線形性をうまく表現できる手法の選択を行う．次に，研究背景で述べた非線形性の考慮を機械学習を用いて行い，先行研究で扱われている一般加法モデル（GAM）の結果と比較する．評価指標としては機械学習で用いられている指標である平均平方二乗誤差（RMSE）, 平均平方二乗誤差率（RMSPE）を用いる．

連絡先: 高橋 佑典，富士通クラウドテクノロジーズ株式会社，東京都中央区銀座 7 丁目 16 番 12 号　G-7 ビルディング，03-6281-5710，yus-takahashi@fujitsu.com

2. 関連研究

2.1 非線形性

[Shimizu 14] では東京 23 区の中古マンションを対象に，線形モデルをベースに置き，物件の平米単価と各説明変数間の関係を非線形モデルとして推計している．具体的には，ノンパラメトリックなモデルである連続量 dummy モデル（DmM）と，AIC を評価指標とした Switching Regression Model（SWR），一般加法モデル（GAM）を用いて中古マンションの取引価格の主要要因が持つ非線形性を明らかにしている．

[小野 02] ではリクルート社の「週刊住宅情報」に掲載された東京 23 区の中古マンションデータを扱い，中古マンション価格に影響を与える要因のうち，「建築後年数」と「最寄り駅までの徒歩時間」に非線形性が存在することを明らかにしている．

2.2 機械学習

[福井 18] では中古マンションの取引データをもつ，レインズの成約データに対してニューラルネットワークを用いて，不動産査定価格を分類問題として解いている．

[大和 18] では不動産情報サイト SUUMO におけるデータを扱い，家賃の推計を行う際に変数の非線形性を表現できる手法として，線形回帰モデルの代わりに決定木をベースとしたランダムフォレストを用いている．

本研究では，関連研究で明らかにされている，住宅価格に影響を与える要因に非線形性が存在する点と，それらの非線形性を機械学習によって表現し，推計価格の精度向上が図れるかを調査する．また，先行研究では主に中古マンションの取引価格データが扱われていることに対し，本研究では戸建住宅の中古取引価格データを扱う．

3. 実験

3.1 データセット

本研究においては，国土交通省が提供している土地総合情報システムのうち，不動産取引価格情報ダウンロードサービス（[土地総合情報システム](http://www.land.mlit.go.jp/webland/download.html)）から取得したデータを用いた．取得したデータは，取引時期が 2005 年第 3 四半期から 2018 年第 3 四半期である東京 23 区内の取引価格情報である．本研究では，戸建住宅の中古取引価格データを扱いたいため，データの種類が「宅地（土地

と建物）」，用途が「住宅」，地域が「住宅地」となっているレコードのみを抽出した．この不動産取引情報には，不動産取引において一般的な取引時期，延床面積，土地面積，最寄り駅までの徒歩分数といった情報が取引価格とともに含まれる．表 1 に本研究で扱うデータの要約統計量を示す．各データはそれぞれ TS:最寄り駅までの徒歩分数，P:成約価格，L:土地面積，S:延床面積，RW:前面道路幅員，BLR:建ぺい率，FRA:容積率，CY:建築年，A:建築後年数，WOOD:木造ダミーである．また，要約統計量で示したデータ以外にも用途地域や市区町村コードといったデータも含まれている．

表 1: データの要約統計量

	count	mean	std	min	25%	50%	75%	max
TS	14424	10.65654	5.411991	1	7	10	14	29
P	14424	48846850	17609200	21000000	35000000	46000000	60000000	99000000
L	14424	93.0633	31.69272	50	70	90	110	200
S	14424	97.72463	24.75509	50	85	95	105	200
RW	14424	4.693795	1.288334	2.5	4	4	5.5	9
BLR	14424	56.56129	7.888145	30	50	60	60	80
FRA	14424	174.0294	76.58877	60	100	150	200	800
CY	14424	1996.819	13.50506	1955	1988	2000	2007	2017
TDY	14424	2011.655	3.736598	2005	2008	2012	2015	2018
A	14424	14.83555	13.07827	1	2	12	23	50
WOOD	14424	0.8840821	0.3201376	0	1	1	1	1

3.2 二段階推計モデル

本研究では，モデルを二段階に分けて推計している．理由として，1 段階目に線形モデルを推計して説明性を担保しつつ，2 段階目のモデルでは多重共線性を持ってしまう要因や非線形性を持つと思われる要因を投入し，1 段階目の誤差を予測し補正することで，モデル全体としての予測精度の向上を図る．

step1.

$$log(y) = \alpha + \beta_1 a_1 + \beta_2 a_2 + \beta_3 a_3 + \cdots + \mu_1$$

step2.

$$\mu_1 = \gamma + \delta_1 b_1 + \delta_2 b_2 + \delta_3 b_3 + \cdots + \mu_2$$

線形モデルを二段階に分けて推計する場合，上記のような 2 つの線形モデルが作られる．本研究では，1 段階目では線形モデルを推計し取引価格を予測する．2 段階目では 1 段階目の予測誤差を目的変数として，非線形なモデルを推計し，取引価格に対する最終的な予測を行う．

本研究では，戸建住宅の取引データに対して一般加法モデル（GAM）を用いて住宅価格に影響を与える主要な要因の非線形性を確認した．一般加法モデル（GAM）を適用した結果を図 1 に示す．

一般加法モデル（GAM）を適用した結果から，極端な傾きの変化は見られなかったものの，S:延床面積，L:土地面積，A:建築後年数，といった主要な要因に対して，非線形性を確認することができた．

3.3 手法の比較

本研究では，線形モデルである OLS をベースラインとし，比較対象として，機械学習の手法である決定木をもとにしたランダムフォレストと xgboost を選択した．また，ニューラルネットワークによる手法として多層パーセプトロンを選択した．以下で各手法についての概要を説明する．

OLS は定数項（切片）と説明変数の係数によって値を予測する線形モデルであり，最小二乗法によって係数と切片を決定する．ランダムフォレストは，決定木とバギングを組み合わせた手法であり，決定木を大量に生成し、各決定木の結果を集計して予測を行う．各決定木は独立して異なる特性を持つように学習する．xgboost は，決定木とブースティングを組み合わせた手法であり，決定木を逐次的に増やしていき、生成済みの決定木の誤差を補正するように、新たな決定木を生成し学習を進めていく．多層パーセプトロンは，入力、中間、出力の 3 層からなるニューラルネットワークの手法。バックプロパゲーションを用いた学習を行う．

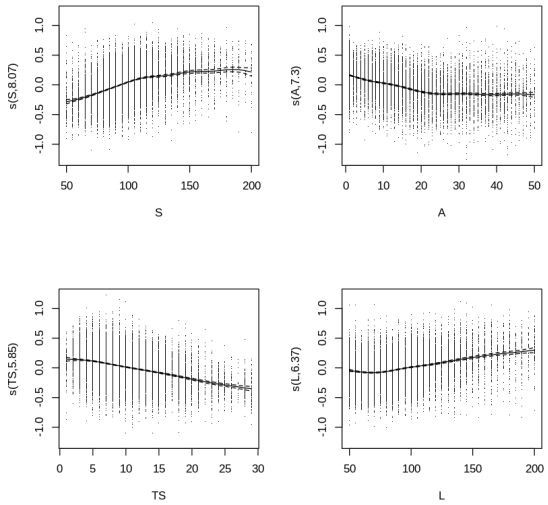

図 1: 各変数と取引価格の関係:GAM

機械学習においては，学習に使ったデータセットだけに過度に適合したパラメータが学習されてしまい，テストデータに対して性能が出ない過学習と呼ばれる状態に陥ることがある．過学習を抑えるために，決定木による手法では特徴量の数や生成する木の深さを設定することが考えられる．また NN による手法では，学習の過程において大きな重みを持つことに対してペナルティを課す Weight decay と呼ばれる正則化の手法が存在する．

本研究における機械学習手法では，グリッドサーチによるパラメータチューニングを行い，過学習を抑制するためにパラメータを定めてからモデルの構築を行った．

3.4 評価

まず，手法間の誤差率分布を比較した結果を図 2 に示す．データは学習データとテストデータを 8:2 でランダムに分割し，各手法において同じ説明変数を投入したモデルで一つ一つの物件に対して取引価格の予測を行い，予測価格/実際の取引価格で定義される誤差率を算出した．その実験を 200 回繰り返し行い，誤差率の平均値の分布を作成した．

結果から，OLS や決定木をベースにした手法は予測結果が大きくはずれてしまう可能性が少ないことに対して，ニューラルネットワークをベースにした多層パーセプトロンでは誤差率の平均値の分布に外れ値が見られた．このことから多層パーセプトロンでは推計結果に大外れが生じてしまうことが確認できた．

3.5 2 段階目投入変数

第 3.4 節の結果から，2 段階目の手法として大外れが少なく，過学習も抑制できていると考えられた xgboost を選択した．2 段階目の機械学習モデルに投入する変数として，他の説明変数への影響が明確でない STATION:最寄り駅ダミーや

CENSUS:大字・通称レベルの地域ダミーを投入してその効果を確認した．また，S:延床面積，L:土地面積，A:建築後年数，TS:最寄り駅までの徒歩分数を説明変数として単独投入し，その効果を確認した．結果を箱ひげ図として図 3 と図 4. に示す．また，各評価指標の値としては表 4. と表 3 に示す．

これらの結果から評価指標を見ると，機械学習における非線形性の考慮が，僅かだが精度向上に寄与していることが確認できた．

4. おわりに

本研究では，戸建住宅データにおける取引価格に影響を与える主要要因に対して，機械学習での非線形性を考慮した 2 段階推計を行った．実験結果より，手法間の比較では，ニューラルネットワークをベースにした手法よりも，決定木をベースにした手法の方が，誤差率のばらつきが少ないことがわかった．また，2 段階推計においては，2 段階目のモデルに個別に説明変数を投入し，その補正結果を調べた．結果は一般加法モデル（GAM）で確認した非線形性を考慮し，精度向上が確認できた．また，合わせて戸建住宅価格に影響を与える最寄り駅ダミーと大字・通称ダミーを二段階目の説明変数として投入し，精度向上を確認できた．今後は，今回得られた手法ごとの誤差率分布をもとに，非線形なモデルを推計する手法ごとの特性を明らかにすることや，モデルに投入する説明変数の数を増やしていき，その効果を検証することが考えられる．

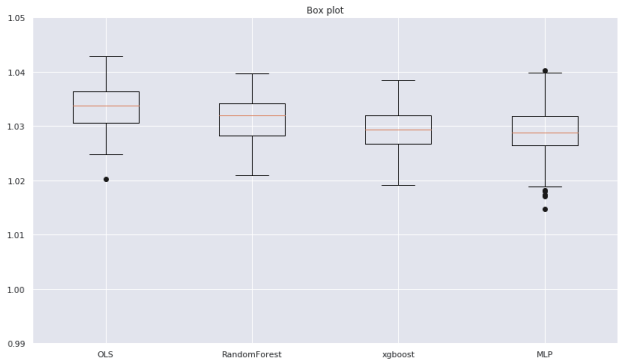

図 2: 誤差率の平均値の分布

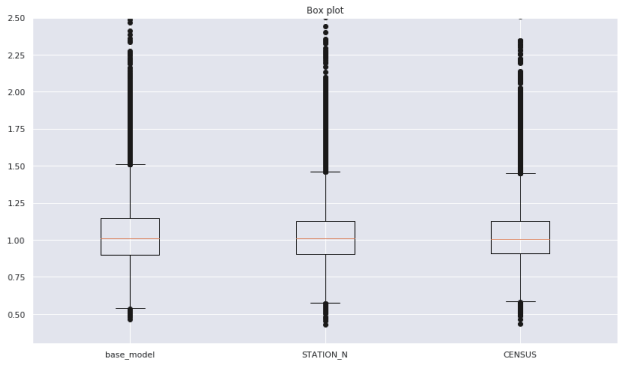

図 3: ダミー変数の効果

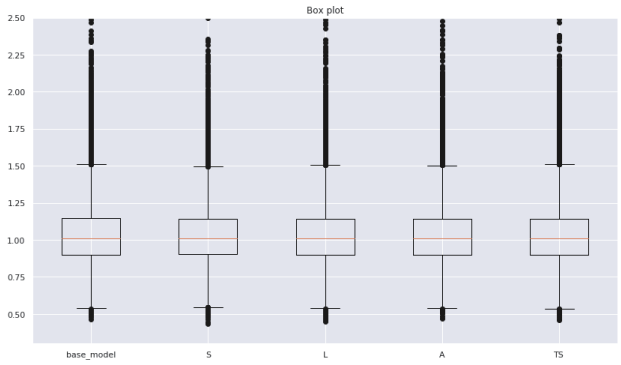

図 4: 説明変数ごとの効果

表 2: 2 段階目投入変数別の効果（ダミー変数）

	RMSE	RMSPE
BaseModel	10038752	5.1895
STATION	9566984	4.7900
CENSUS	9364350	4.6412

表 3: 2 段階目投入変数別の効果（連続量）

	RMSE	RMSPE
BaseModel	10038752	5.1895
S	9942438	5.1025
L	9976350	5.1309
A	9902803	5.1253
TS	10033856	5.1733

参考文献

[Shimizu 14] Shimizu, C., Karato, K. and Nishimura, K.: Nonlinearity of housing price structure assessment of three approaches to nonlinearity in the previously owned condominium market of Tokyo, Int. J. Housing Markets and Analysis, Vol. 7, No. 4, pp. 459-488（2014）.

[小野 02] 小野宏哉，高辻秀興，清水千弘：「品質を考慮した中古マンション価格モデルの推定」，麗澤経済研究，第 10 巻第 2 号，pp.81-102，2002.

[福井 18] 福井光，阪井一仁，南村忠敬，三尾順一，木下明弘，田司郎：「レインズのニューラルネットワークを用いた不動産価格査定について」，The 32nd Annual Conference of the Japanese Society for Artificial Intelligence, 2018.

[大和 18] 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／第 31 巻第 1 号, pp.78-83, 2017.