The 34th Annual Conference of the Japanese Society for Artificial Intelligence, 2020

賃貸住宅入居希望者からの問い合わせがあるかどうか，募集物件情報から予測できるか

～解析手法による予測精度の違いとビジネスへの適用のしやすさに関する考察～

Predicting inquiry from potential renters using property listing information - Prediction accuracy and applicability to business -

宗健[*1]
Takeshi SO

新井優太[*2]
Yuta Arai

[*1] 大東建託賃貸未来研究所
Daito Trust Construction Co.,Ltd

[*2] 麗澤大学
Reitaku University

In this study, We deduced how accurate the number of inquiries from potential tenants for housing available for rent can be predicted based on the attributes of the housing, using multiple statistical methods, and compared the results. The purpose of this study is to show these results as case studies.

Confusion matrices were calculated based on the results deduced with three methods – the classical logistic regression, RandomForest, and XGBoost – and prediction accuracies were verified. The results showed that the accuracy of XGBoost was the highest, followed by that of logistic regression.

It is sometimes desirable to use logistic regression, which is easy to interpret from the perspective of application to business, because the differences in accuracy among the statistical methods are not significant. It is thus important in business to take into account the accuracy, ease of interpretation, and research structure and select the most appropriate statistical method.

1. 研究の背景と目的

データ解析に用いられる統計手法は日々進化しており，多くの研究者が新たなアルゴリズムの開発や，統計ソフトやコンピュータ言語への関数としての実装に取り組んでいる．

一方で，実装される関数が増えれば増えるほど，実務的にはどの関数をどのように使えば良いか悩むことも多くなってきている．解決すべき課題の目的によって離散変数の推定，判別，変数の集約などに大まかには分けられるとは言え，具体的に適用される関数の選択については，分析者の得意不得意や経験によって決められる場合も多く，古典的な手法よりも新しい統計手法を用いる方が，ビジネス領域においてはクライアントの反応が良いという風潮もある．

その原因として，一つの解決すべき課題について，複数の関数を用いて結果を比較し，どのような場合にどのような関数を用いるかという実務的なケーススタディが不足していることが背景として考えられる．多くの応用研究では，課題に対するデータと適用された手法（関数）および結果だけが示されている場合がほとんどであり，なぜその手法を用いたのか，他の手法を用いた場合と結果がどのように異なるのかが示されることは少ない．

このような背景から，本研究では実務的なケーススタディとして，「賃貸住宅入居希望者からの問い合わせがあるかどうかについて，募集物件情報から予測できるか」という課題について，Logistic 回帰・XGBoost ・RandomForest の 3 つの手法を用いて推定を行い，精度の違いやビジネスへの適用のしやすさ等について考察を加え，ケーススタディとして示すことを目的としている．

こうしたケーススタディを積み重ねていくことは，実務的に分析を行う際に，検証課題と状況，技術者のスキルや経験，人数といった分析体制によって適切な手法を選択することに貢献できると考えられる．

連絡先：宗健，大東建託賃貸未来研究所，108-8211 東京都港区港南 2-16-1，st102456@kentaku.co.jp

2. 先行研究

本研究で課題とした「賃貸住宅入居希望者からの問い合わせがあるかどうかについて，募集物件情報から予測できるか」については，賃貸空物件が入居されるまでの期間に関する研究として渡邊・一藤・鈴木・山下(2019)がある [1)].

しかし，不動産領域以外では，例えばホテルや旅館等の宿泊施設の予約確率については，青田・田中(2006)のような予測モデルの研究が様々な企業で行われており，単価と予約率から売り上げを極大化する，レベニューマネジメントが進化している．それは宿泊については，世界的にも大規模な企業が多く，分析に必要なデータが揃いやすいことが背景にある．一方，不動産領域では国によって条件や状況が大きく異なり，日本国内では分析に耐えられるだけの大量のデータを保有している事業者がほとんどいないことが，研究の行われていない大きな理由だと考えられる．

不動産領域での分析では，家賃や中古マンション価格の推計が多く行われており，山崎・浅田・瀬下・清水(2007)，宗・新井(2018)，宗(2018)，So・Shimizu(2019)といったものがある．しかしこれらの先行研究では，古典的な重回帰分析が用いられているだけで，複数の統計手法の比較は行われていない [2)].

複数の手法を比較した不動産領域での先行研究では，清水(2017)があり，伝統的な回帰分析とニューラルネットワーク，回帰木の誤差率の比較を行い，非線形モデルと線型モデルについての考察など示唆に富む結果を提示している．

しかし，こうした先行研究では，ビジネスへの適用のしやすさといった実務的な観点での考察は不足している．

3. 研究方法

3.1 データ

本研究では，賃貸不動産仲介会社 [3)]が 1 都 3 県（埼玉県・千葉県・東京都・神奈川県）で募集を行っている賃貸物件を対象として分析を行う．賃貸住宅市場においては，1 月～3 月が繁忙

期とされており，特に内覧などの問い合わせが増える．そのため今回の研究では繁忙期である 2019 年 3 月と閑散期である 2019 年 6 月の 2 時点について，どのような物件属性であればメールよる問い合わせ（反響）がはいるのか，について分析を行う．この際，物件属性の不均一性を低減するために，分析対象を面積 15 ㎡～30 ㎡のシングル向け物件に限定した．

表 1 は使用する物件データの記述統計量である．

表 1. 物件データの記述統計量

		家賃(万円)	面積	築後年	駅徒歩(分)
2019年3月	物件数	34,736			
	平均値	8.82	24.02	8.70	7.06
	標準偏差	2.86	3.26	8.42	3.59
	最小値	2.00	15.00	0.00	1.00
	最大値	22.60	30.00	30.00	15.00
2019年6月	物件数	15,641			
	平均値	8.80	24.27	8.60	6.57
	標準偏差	2.75	3.23	8.00	3.37
	最小値	2.50	15.00	0.00	0.00
	最大値	22.60	30.00	30.00	15.00

対象のエリアにおいて各時点の 1 か月間で募集のあった物件数はそれぞれ 2019 年 3 月で約 3.5 万件，2019 年 6 月で 1.6 万件である．分析対象は，異常値を除去するために家賃は 2 万円～30 万円，建築後年数は 30 年以下，駅徒歩 15 分以内としている．

表 2 は反響が入った物件の記述統計量である．

表 2. 反響が入った物件の記述統計量

		家賃(万円)	面積	築後年	駅徒歩(分)
2019年3月	物件数	682(反響率:1.96%)			
	平均値	7.02	23.93	10.28	7.19
	標準偏差	2.23	3.66	9.12	3.58
	最小値	2.60	15.01	0.00	1.00
	最大値	19.80	30.00	30.00	15.00
2019年6月	物件数	442:(反響率:2.82%)			
	平均値	7.90	23.73	8.31	6.99
	標準偏差	2.45	3.60	7.97	3.53
	最小値	2.58	15.65	0.00	1.00
	最大値	16.40	30.00	30.00	15.00

表 2 と表 1 を比較すると，反響が入る物件は全募集物件のうち 2%程度のごく一部の物件に限られており，大多数の募集物件には反響が入っていないとわかる [4]．

記述統計量に注目してみると，反響物件のほうが家賃は 2 万円ほど安く，築後年数が 1.5 年ほど古いものの，それ以外の面積・駅徒歩で大きな差はみられない．このことから，物件に対する反響有無を面積などの基本的な項目から予測することが困難であることがわかる．

そこで本研究では，ロジスティック回帰分析・XGBoost ・RandomForest の 3 つの手法について，共通の説明変数を用いて反響の予測を試み，精度の比較を行う．

3.2 使用変数

表 3 は用いた説明変数である．使用する変数には，記述統計量で示した面積，築後年，駅徒歩に加えて設備，敷金，礼金などの一時金情報等を用いる．物件種類や構造などの物件属性に関するものはカテゴリカル変数，設備に関するものは 0,1 のダミー変数である．また，面積や築後年等の離散変数についてはダミー変数化して分析をおこなった [5]．

このほか，特殊な変数として，相場家賃からの乖離を用いている．これは分析対象データから算出した推計家賃と，実際の募集家賃との乖離を表したものである．厳密には，説明変数の中に推計値を入れることは好ましくないが，近年のようにインターネットサイトで多数の物件を消費者が比較できる場合，消費者にも物件の割高・割安が感覚的に認知されていると考えられる．そのため，相場家賃よりも割安な物件のほうが反響は入りやすいと考えられるため，今回は説明変数の 1 つとして採用した．

表 3. 説明変数一覧

変数の属性	変数名
物件属性変数	物件種類 構造
量的変数	面積 駅徒歩 築後年
設備系変数	エレベーター 宅配ボックス オートロック 温水便座 間取り 1階
一時金変数	礼金区分 敷金区分
住所	市区郡
独自変数	市場家賃からの乖離率

3.3 検証手順

2019 年 3 月の繁忙期と 2019 年 6 月の閑散期で反響の入る物件に差がある可能性もあるため．2019 年 3 月データと 2019 年 6 月データでそれぞれ分けて推計を行う．

検証は，反響有無でデータを 8:2 に分割し，8 を学習用データ，2 をテスト用データとする．さらに，今回は反響有データが著しく少ないため，学習用データに対してはアンダーサンプリングを行ったうえで学習に使用し，テスト用データについて反響確率を計算する．この試行を各手法について 100 回行い，物件ごとに得られた反響確率を用いて Confusion Matrix を計算し，評価を行う．

4. 結果

図1は，2019 年 3 月について，実際に反響が入った物件に対して推計された反響確率の分布である．

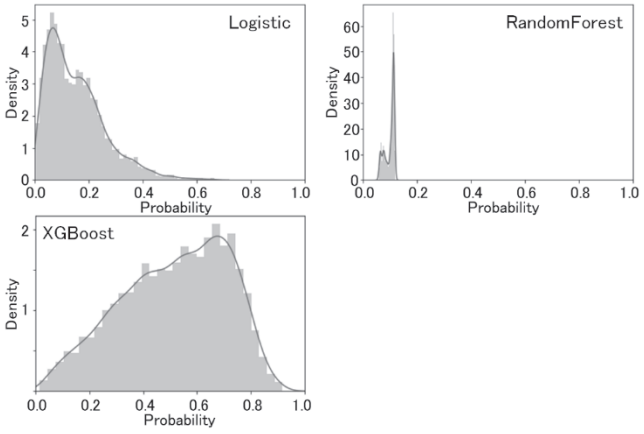

図 1. 推計された反響確率の分布

図 1 では，手法によって推計された反響確率の分布が大きく異なることが示されている．例えば，中央値をみてみると，Logistic 回帰では 0.14，XGBoost では 0.58，Random Forest では 0.12 と異なっている．これだけ分布の形状が異なると，反響確率について共通の閾値を設定することは適切ではない．

そこで本研究では，実際に反響が入った物件に対する反響確率の中央値を閾値として採用し，表 4 のような Confusion Matrix を計算し，推計精度の検証を行った．

表 4. Confusion Matrix

	手法		推計反響なし	推計反響あり
2019年3月	Logistic	反響なし	77.00%	21.07%
	中央値:0.13833	反響あり	0.97%	0.97%
	XGBoost	反響なし	**78.56%**	19.51%
	中央値:0.57917	反響あり	0.97%	0.97%
	RandomForest	反響なし	67.40%	30.67%
	中央値:0.11509	反響あり	0.97%	0.97%
2019年6月	Logistic	反響なし	**74.63%**	22.56%
	中央値:0.13114	反響あり	1.41%	1.41%
	XGBoost	反響なし	73.37%	23.81%
	中央値:0.536389	反響あり	1.41%	1.41%
	RandomForest	反響なし	63.36%	33.83%
	中央値:0.10777	反響あり	1.41%	1.41%

実際には反響なしで推計上も反響なしの検知率[6)]に注目すると Logistic 回帰と XGBoost が同程度であるとわかる．2019 年 3 月では XGBoost のほうが高い検知率を示しているものの，2019 年 6 月では Logistic 回帰のほうが高い検知率を示している．実務上は募集件数が多く，成約数も多い 2019 年 3 月の繁忙期のほうが重要であり，その意味では XGBoost のほうが Logistic 回帰よりもやや精度が高いと言える．

一般的には，このような分類問題において，Random Forest や XGBoost は高精度で分類可能であることが知られているが，今回のように閾値を調節することで，旧来の手法である Logistic 回帰も実務的に十分な分類精度を示すと言える．

5. 考察および今後の課題

本研究では，一つの課題について Logistic 回帰・XGBoost・RandomForest の 3 つの手法による結果を比較したが，今回のケースでは，実際のビジネスに適用する手法としては，古典的な logistic 回帰を用いるほうが望ましいと考えられる．

それは，logistic 回帰と XGBoost がほぼ同程度の予測精度である場合，logistic 回帰のほうが結果を解釈しやすく，推計結果を利用するビジネスの当事者が，判別の理由を理解しやすいからである．

ビジネスにおける統計手法の利用では，説明力は必要なく精度が優先される場合もあるが，精度が同程度であれば，解釈しやすくロジックを理解しやすいほうが，実際の利用場面での評価や性能向上のための意見収集が容易である．そしてそれがビジネスでの利用を定着させ，モデルの精緻化にも繋げやすいと考えられる．

また，こうしたモデル開発に取り組む際には，最新の統計手法を常に把握しいつでも使えるようにしておくことも重要だが，最新の手法だからといって優先的に利用することは必ずしも正解にはならない．ビジネスへの適用は，複数の手法を試し，精度と説明力のバランスを考慮して使用する統計手法を選択すべきである．また，統計手法によっては過学習が起こりやすいことにも留意が必要である．

今後の課題としては，データを適切に層別化した際に，用いる統計手法によって結果がどのように異なるかを検証することがある．層別化のやり方によって，結果が異なることも想定されるためであり，統計手法によって適切な層別化の手法が異なる可能性もあるからである．

謝辞

本研究の分析のうち XGBoost および RandomForest のパラメータチューニングや分析について，富士通クラウドテクノロジーズ株式会社の堀貴仁氏に協力して頂いた．ここに記して感謝の意を表する．

参考文献

[渡邊 2019]渡邊 隼史，一藤 裕，鈴木 雅人，山下 智志:Web 不動産データを用いた空物件が入居されるまでの期間に関するデータ特性を考慮した統計モデリング，人工知能学会全国大会論文集 JSAI2019(0), 1D3OS10b03-1D3OS10b03, 2019

[青田 2006]青田 良紀，田中 康秀:オンライン取引による価格決定要因および価格差の分析 : ヘドニツク・アプローチによるホテル客室料金を対象として，生活経済学研究,2006 年 22.23 巻 71-79

[山崎 2007]山崎 福寿，浅田 義久，瀬下 博之，清水 千弘:住宅資本コストが住宅所有形態に及ぼす影響についての実証分析，住宅総合研究財団研究論文集 33(0), 335-345, 2007

[宗 2018]宗健，新井 優太:富裕層および団地の集積が家賃に与える影響，都市住宅学 2018(103), 126-131, 2018

[宗 2018]宗健:地域の共同住宅空室率が中古マンション価格に与える影響，日本不動産学会誌 32(1), 106-113, 2018

[SO 2019]So, T and C.Shimizu (2019):Housing Facilities and Housing Rent,Purchasing and Supply Management(Peer review book), ISBN 978-1-78984-973-8

[清水 2017]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 31(1), 45-51, 2017

脚注

1) 賃貸物件の入居までのプロセスでは，入居希望者が自ら検索して物件に問い合わせるまでと，実際に不動産会社を訪問して詳細情報の説明を受け物件を内見して入居を決定するまでのプロセスに大きく分かれる．後者のプロセスでは広告されている情報以外の要素が大きく，物件情報だけを使ったモデル構築には説明変数が不足していると考えられる．

2) 民間のサービスとしては，不動産ポータルサイトの SUUMO・LIFULL HOME'S・athome はいずれも家賃相場を誰でも使える形で提供している．LIFULL HOME'S は PriceMap という形で中古マンション価格の推計値をマップで閲覧できるサービスを提供している．また，不動産 tech ベンチャーである株式会社ターミナルは地域限定ではあるが商用サービスとして家賃査定サービス(スマサテ)を提供している．また，SRE 不動産(旧ソニー不動産)も，中古マンションの価格査定サービスを提供している．しかし，これらのサービスはいずれもどのようなデータをどのような手法でモデル化したのかという情報が十分に公開されているわけではない．

3) 募集されている物件は特定の会社のものではなく，いわゆる先物物件と呼ばれる他社管理の物件が多数含まれている．

4) 実際の成約は特定のウェブサイトからのメール反響だけで決まるものではなく，電話問い合わせや直接来店，他社からの紹介など多くのルートがある．また，同じ物件を複数の不動産会社が同時に募集することから，物件あたりの反響数はかなり低くなる傾向にある．

5) 賃貸住宅の築後年や面積については，家賃に対して非線形の影響を及ぼすことが知られているが，今回の分析ではこれらの離散変数をダミー化しているため，線形としては扱われていない．

6) 実務的には，反響が入る可能性が高い物件を優先的に募集することで，業務効率を向上させることができるというメリットが大きい．