アンサンブル学習とニューラルネットワーク”じゃらん”データを使った分析

長谷川，博康
羽森，茂之

(Citation)
國民經濟雜誌,220(3):17-29

(Issue Date)
2019-09-10

(Resource Type)
departmental bulletin paper

(Version)
Version of Record

(JaLCDOI)
https://doi.org/10.24546/E0041859

(URL)
https://hdl.handle.net/20.500.14094/E0041859

アンサンブル学習とニューラルネットワーク
“じゃらん”データを使った分析

長谷川　博康
羽森　茂之

国民経済雑誌　第 220 巻　第 3 号　抜刷
2019 年 9 月

『国民経済雑誌』第220巻第3号（2019年9月），17-29

アンサンブル学習とニューラルネットワーク “じゃらん”データを使った分析*

長谷川　博康[a]
羽森茂之[b]

AI（人工知能）やマシンラーニング（機械学習）が，金融業や製造業，小売業，マーケティング分野など様々な業界，分野において取り入れられている。本論文の目的は，決定木，アンサンブル学習，及びニューラルネットワークを旅行サイト“じゃらん”のデータにあてはめ，各分析手法の予測性能に関して比較検討を行うことである。本論文で得られた主要な結果は，以下のとおりである。1）決定木やニューラルネットワークに対して，アンサンブル学習 XGBoost が有効である。2）ランダムフォレストの正答率は，学習データに対しては高いが，検証用データでは正答率が劣る。3）C5.0 とニューラルネットワークの正答率を比較すると，学習用と検証用データとも C5.0 が高い結果を得た。しかし，モデル評価の指標である AUC 値や F 値に関しては学習用と検証用のデータともにニューラルネットワークが高い値が得られた。4）従来の決定木の手法である CHAID や CART についても正答率やモデル評価の指標において高い値をとることが明らかになった。

キーワード　機械学習，アンサンブル学習，ニューラルネットワーク，じゃらん

1　はじめに

最近のニュースや新聞記事では，AI（Artificial Intelligence：人工知能）や機械学習について多く取り上げられている。このことは，一般に AI や機械学習の認知が広まり，社会やビジネスにおいても使用されていることを示している。これら一般に広まったきっかけは，Deep Blue（IBM が開発したチェス専用のスーパーコンピュータ）が世界チャンピオンに勝利したことである。その後，コンピュータ囲碁の AlphaGo やコンピュータ将棋の Ponanza が，次々とプロ棋士に勝利したことにより，更に認知度が高まった。そして，経済界や産業

a　神戸大学大学院経済学研究科，brandon_90240@hotmail.com（責任著者）
b　神戸大学大学院経済学研究科，hamori@econ.kobe-u.ac.jp

界においても，AIや機械学習の導入へと次々と進められてきている。例えば，金融界では人工知能による与信モデルや不正検知（Fraud）などがあり，マーケティングの業界では商品のリコメンドなどで使用されている。また製造業では欠品率を下げるための予測精度の改善や画像認知機能による判断など，AIや機械学習の導入が様々な分野で取り入れられている。これら予測技術や予測モデルは最近になって開発されたものではなく，以前から存在するニューラルネットワークや決定木の手法など統計解析や数学的手法を使ったモデルである。このような予測モデルが近年になって改良され，理論的なモデルから実用的なモデルへと改良され，実社会においても使用されるようになったのである。

このような背景を元に，本論文ではニューラルネットワークや決定木等の予測モデルをあてはめ，予測精度を比較検討し，その有効性を確認することである。また，本論文の特徴は，旅行サイト“じゃらん”のデータを使ってモデル精度の比較を行い，そのモデルの有効性について分析したことである。その結果，XGBoost（eXtreme Gradient Boosting）が有効[1)]であることが明らかになった。また，ランダムフォレストは，学習用データでは正答率は高いものの検証用データでは正答率が低下した。その点でも，XGBoostが支持できる。また，C5.0の正答率は高いものの，モデル精度の評価であるAUC値やF値が劣っている。これに対してニューラルネットワークでは，正答率はC5.0よりも良くないものの，AUC（Area Under the Curve）値やF値において高い値であった。また，従来の決定木の手法であるCHAID（Chi-squared Automatic Interaction Detection）やCART（Classification and Regression Tree）についても，予測結果の正答率やモデル評価の指標においても大きくは劣らないことが明らかになった。

本論文の構成は，以下の通りである。第2節では，決定木とニューラルネットワークを説明するともに，既存の予測モデルについての学術研究について述べる。第3節では，使用データとモデル精度の評価について述べる。第4節では，結果の正答率とモデル評価の指標の結果を考察する。第5節は，本論文のまとめである。

2　予測モデルの概要と既存研究[2)]

2.1　決定木のモデル

本節では，予測モデルについて説明する。決定木モデルは，カテゴリ目標変数（従属変数）に対して，説明変数（独立変数）で順次分岐を行い，枝葉に分かれる。その枝葉に分かれたグループ（セグメント）をノードといい，これ以上成長しなくなったノードを最終のノード（ターミナルノード）という。たとえば，商品の購入（する・しない）を予測する場合，商品の購入にもっとも関連する説明変数を選択し，その変数で分岐を行い，樹木を成長させるのが決定木である。

最初に提唱された決定木の手法は，CHAID（Chi-squared Automatic Interaction Detection）である。CHAID は，Kass（1980）によって提唱され，Magidson（1993）によって開発された決定木の手法である。CHAID は，カイ 2 乗検定である独立性の検定を使用する決定木の手法である。この手法は，設定したカテゴリ目的変数とそれに関連するであろうとする複数のカテゴリ説明変数間で独立性の検定を実行する。その結果，説明変数間で p 値を比較し，その中で最も小さい p 値の説明変数で分岐するのである。目的変数と説明変数間において，独立性が低い説明変数，つまり最も関係がある判断し，分岐を行う。また，その分岐されたそれぞれのセグメントにおいて独立性の検定を行い，関連が高いと思われる説明変数で分岐する。このように順次関連の高い説明変数で分岐することで樹木を成長させる。決定木のモデルが作成されれば，ツリーは複数のセグメントに分割される。そして，その最終の分割されたセグメントのことをターミナルノードと呼ぶ。ターミナルノードでは，目的変数のカテゴリが比較され，その最頻値に予測する。このように，CHAID の手法では，独立性の検定を使用して，ツリーが分岐され，樹木が成長する。そして，ターミナルノードの最頻値に予測するモデルが決定木による予測モデルとなる。

次に開発された決定木の手法は，CART（Classification and Regression Tree）で，Breiman（1984）によって開発された手法である。分類と回帰の樹木とも呼ばれ，目的変数がカテゴリ型の場合は分類の樹木で，目的変数が連続型の場合には，回帰の樹木となる。CART は，CHAID の手法と同様，目的変数と関係が高い説明変数を選択し，順次分岐を行う。分岐の際に使用するのは，Gini の分散の測度を使用する。この Gini の分散の測度から分岐した際の各セグメントにおける不純度を計算する。この不純度は，分岐された各カテゴリにおいて，目的変数の偏りの測度である。この偏りである不純度が分岐前のセグメントと比較して最も改善される説明変数を使用して樹木を成長させる。つまり，Gini の分散の測度からセグメントごとの不純度を計算し，分岐前と分岐後の不純度の比較を行って，その改善度から分岐する変数を選択する。不純度とは，各変数のカテゴリのばらつき度合いを示すもので，完全に純粋なノードは不純度が 0 となる。これは，目的変数のカテゴリに集中する度合いである。この不純度の測度が，Gini の分散の測度である。この分岐により作成されるそれぞれの子ノードが，親ノードよりも純粋であるように分岐し，ここで純粋なノードへとなるように樹木を成長させる。つまり，目的変数のカテゴリ値が 1 つになるように説明変数を選択することで，なるべく予測の精度を高めるように樹木を成長させる。しかし，ツリーを成長させ過ぎるとモデルの再現性がないことから，一旦ツリーを成長させ，その後ツリーの剪定を行うことで，安定的なツリーが作成されることを Beriman（1984）が発見し，CART 以降の決定木においては，剪定が行われている。

C 5.0 の手法は，エントロピーの測度である情報量基準をもとにした分岐の手法である。

この手法は，Quinlan（1993）によるものである。エントロピーの測度は，不確実性の測度とも呼ばれ，情報の不確かさを表す散らばりの測度である。エントロピーの測度をもとにした情報利得を計算し変数を選択するのであるが，この情報利得とは分岐前の平均情報量と分岐に必要な情報量の差から計算されるものである。この差が情報利得となり，この差が最大になる変数を選択し樹木を分岐させる。また，C5.0 ではこの情報利得から潜在的な情報量を使って正規化した利得比基準を使用している。各説明変数のデータ型によって情報量が変わることからデータの型の影響を取り除いたものである。この利得比基準によって選択された説明変数で分岐し，樹木を成長させるのが，C5.0 の決定木である。また，C5.0 も CART 同様，オーバーフィットを避けるため剪定が行われる。

2.2　アンサンブル学習

アンサンブル学習は，複数の樹木を成長させる手法である。そのアンサンブル学習の一つに，ランダムフォレストの手法がある。ランダムフォレストは，複数の CART の樹木からなる手法である。この手法は CART の開発者である Breiman（2001）によるものである。また，アンサンブル学習にブースティングもある。ブースティングは，逐次的に重みを変化させながら異なる学習器を作る手法で，複数のモデルを組み合わせて高い精度の学習器を作る手法である。この代表的なアルゴリズムとして，AdaBoost があり，この手法は Freund and Schapire（1997）により開発された手法である。この学習アルゴリズムは弱仮説から誤り率をできるだけ少なくすることでモデルを強化する手法である。これらアンサンブルの手法からランダムフォレストでは，複数樹木の成長から結果を結合するモデルで，予測結果を多数決することによって決定する方法である。ランダムフォレストの開発者は，Breiman（2001）である。Breiman（2001）では，ランダムフォレストについて，Adaboost よりも精度は良く優れており，また，バギングやブースティングよりも速度が速いと述べている。

次に XGBoost（eXtreme Gradient Boosting）は，Gradient Boosting（勾配ブースティング）とランダムフォレストを組み合わせたアンサンブル学習である。この XGBoost は，Chen and Guestrin（2016）によるものである。勾配ブースティングを使うということで，ブースティングに勾配を付けたモデルを作成する。ここでの勾配は，ブースティングによるアンサンブル学習なので前回までのモデルに対しての勾配を付けるということである。誤差を小さくするために前回のモデルを使用した変数に対して最適化を行い目的関数の勾配として取り出すのである。また，過学習によるデータへのオーバーフィッティングを避けるためにペナルティ項である正則化項を設けている。そしてこの正則化項を含めた最適化が行われるのである。このように，ランダムフォレストを使うのであるが，この勾配によりアンサンブル学習が強化される。また，正則化項により，オーバーフィットを避けることができる点がラン

ダムフォレストとの違いである。

2.3　ニューラルネットワーク

ニューラルネットワークは，神経の伝達関数としてモデル化されたものである。これは，生物の神経系は多数のニューロン（神経細胞）で構成され，それぞれが結合し並列処理を行うことをモデル化したものである。そのニューロンは，樹状突起，軸索，細胞体から成り立っており，各ニューロンの樹状突起は，シナプスを通して他のニューロンからの入力信号を受け取るのである。これは McCulloch and Pitts（1943）が考案した生物系内のニューロンの動作原理に基づいたものとしてモデル化され，当初は単位ステップ関数の活性化関数が使用されていた。この McCulloch and Pitts（1943）を元に，Rosenblatt（1958）はパーセプトロンを考え，3 層の階層構造パーセプトロンへと発展させた。しかし様々な欠点があり，その後 Rumelhart（1986）らによってバックプロパゲーション（BP：Back Propagation 逆誤差伝播法）が提案された。これよって，階層構造ニューラルネットが関心を集めるようになったのであるが，各階層間の重みは，活性化関数のシグモイド関数を使用してバックプロパゲーション法により調整される。このバックプロパゲーションは，入力パターンに対する誤差関数，および全ての入力パターンに対する誤差である総誤差関数を最小化する最小化問題を取り扱うものである。また，この 3 層構造を持つニューラルネットワークは，入力層と中間層（隠れ層）と出力層を持ち，マルチレイヤーパーセプトロン（MLP：Multilayer Perceptron）と呼ばれるニューラルネットワークの 1 つである。この中間層を深く幾重にもしたものが，ディープニューラルネットワークと呼ばれる。

2.4　既存の予測モデル研究

本論文では，決定木からアンサンブル法を使った決定木とニューラルネットワークを作成し，その比較を行うものであるが，このような予測モデルの手法を比較した論文は，様々な分野で研究が行われている。Razi and Athappilly（2005）では，喫煙者に対して病気の日数を予測させるモデルで比較している。そこで使用されている予測モデルの手法は，非線形回帰，CART，ニューラルネットワークでその予測誤差を使用してモデル比較を行っている。結果，ニューラルネットワークと決定木モデルの方が，非線形回帰よりも優れた結果を示している。また，ニューラルネットワークに対して，CART の手法がより大きな問題へと拡張可能であり，少ないデータに対しても扱うことができると指摘している。他にも医学の分野では，Kurt, Ture and Kurum（2008）が冠動脈疾患（CAD：Coronary Artery Disease）の予測を行っており，そこではロジスティック回帰（LR：Logistic Regression），CART，ニューラルネットワークのマルチレイヤーパーセプトロン（MLP）と放射基底関数（RBF：Radial Basis

Function），自己組織化マップ（SOFM：Self -Organizing Feature Maps）の手法で比較を行っている。モデルの結果，ROC（Receiver Operating Characteristic）曲線やAUC（Area Under the Curve）の比較，有意差検定，階層クラスタ（HCA：Hierarchical Cluster Analysis），多次元尺度法（MDS：Multi-Dimensional Scaling）からモデル比較を行っている。結果，ロジスティック回帰（LR），CART，ニューラルネットワークのマルチレイヤーパーセプトロン（MLP），放射基底関数（RBF）の手法の方が自己組織化マップ（SOFM）よりも優れていることを示している。また，他にもAustin, Tu, Ho, Levy and Lee（2013）では，心不全患者の分類について予測モデルを使って予測判別を行っている。ロジスティック回帰とCARTの比較をおこなっているが，バギング，ブースティグ，ランダムフォレスト，SVM（Support Vector Machine）についても言及しており，モデルの精度を比較している。バギングやブースティングを使用した柔軟な決定木は，モデルの精度の改善をもたらすことを述べている。

その他の分野では，Tso and Yau（2007）が電力エネルギーの消費研究における予測モデルの比較を行っている。従来の回帰分析から，決定木とニューラルネットワークを扱っている。正答率については予測誤差から比較され，決定木は回帰分析の手法やニューラルネットワークよりもよりシンプルで正確に予測される指摘されている。また，Ahmad, Mourshed and Rezgui（2017）の論文では，エネルギー予測のモデルを比較している。そこで，ニューラルネットワークとランダムフォレストの比較が行われている。ホテルの電力消費量を予測するモデルで，2つのモデルの比較を行った結果を示しており，ニューラルネットワークは，ランダムフォレストよりもわずかに優れていることを示している。しかし，両方のモデルとも精度の上では学習用データと検証用データにおいても優れたパフォーマンスの結果であると述べている[3)]。

以上のように，様々な分野で予測モデルが予測可能であるかどうか，モデルの精度が高いかなどを比較している。本論文でも決定木からアンサンブルのモデル，ニューラルネットワークを比較する。

3　旅行サイト“じゃらん”データを使った予測モデルの分析

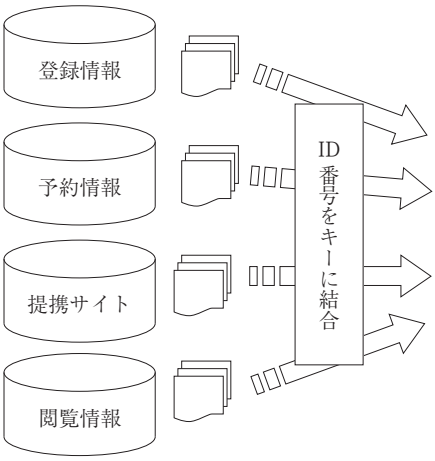
ここで使用するデータは，旅行サイト“じゃらん”の予約情報のデータである。“じゃらん”は，株式会社リクルートライフスタイルの旅行サイトである。日本最大級の宿・ホテルなど旅行に関する予約から，ゴルフ，遊び・体験などを扱っている。株式会社リクルートライフスタイルは，旅行以外にも，飲食や美容，ヘルスケアなどにも事業を展開している。この“じゃらん”の予約情報のデータを元に，予測モデルの比較を行った。

データは，特定期間の新規会員を対象とし，申し込みから1年以内に1回以上利用した会員を対象にする。今回の分析では，登録後利用しない場合や登録後1年以上経過してから利

用した場合には対象外となる。また利用期間は，1回目の利用から18カ月以内とし，この期間内に2回以上利用するかどうかを分析の対象にした。つまり，これを目的変数とした予測モデルを作成した。予測モデルの説明変数は，年齢，性別などのデモグラフィックデータ，予約状況，利用金額，回数，宿情報，時間帯や曜日などである。[4)] 他の提携サイトからの情報やサイトの閲覧情報なども利用した。以上の今回使用した説明変数は，72変数である。これらを元に，予測モデルを作成した。

各データは，それぞれのデータベースに保存されているので，これらデータをマージして1つのファイルに結合する必要がある。以下の図に示すのがデータの結合イメージである。様々なデータベースからIDをキーにして結合し，分析用のデータファイルを作成する。図の右側に示すものが分析用データファイルのイメージである。

図1　データ結合のイメージ：分析用データファイルの作成



ID	Y	X01	X02	X03	X04	X05	X06	X07	X08	X09
1001	1	1	19	9	3	120	2	5	4	3
1002	1	2	20	7	1	100	1	5	4	2
1003	0	2	19	7	3	45	2	4	2	4
1004	0	2	20	7	3	20	2	4	3	3
1005	1	1	21	7	2	70	1	4	4	2
1006	0	1	22	10	3	80	2	1	2	3
1007	1	1	20	7	3	60	7	1	1	1
1008	1	1	24	7	3	100	3	2	1	4
1009	0	2	20	7	3	50	3	3	1	1
1010	0	2	22	9	3	70	3	4	2	2
1011	1	2	23	8	1	80	4	1	1	1
1012	1	2	22	7	3	70	6	1	1	1
1013	1	2	22	7	3	70	3	4	2	2
1014	1	2	23	9	3	50	3	3	1	1

ここで使用するデータは，予測の対象カテゴリで再利用を“する・しない”を均等に分割し，サンプリングを行った。対象のデータは，10万件のデータを用意した。従って，データは予測対象のカテゴリ（再利用する・再利用しない）について，それぞれ5万件ずつ用意し，計10万件の対象データである。また，予測モデルを作るにあたってデータの分割を行った。モデル精度の比較のため，学習用データと検証用データで3つのパターンで分割を行った。ここで分割した割合は，5：5，7：3，9：1の3つの方法で分割した。データ分割の割合が5：5の場合，学習用データ5万件でモデル作成を行い，残りの検証用データ5万件で検証作業を行い，予測モデルの正答率を比較した。同様にして，7：3の場合，7万件のデータでモデル作成を行い，3万件のデータで検証作業を行う。また，9：1の場合には，9万件のデータでモデル作成を行い，1万件のデータで検証作業を行った。

モデルの精度の指標として，AUC（Area Under the Curve）とF値を使用する。このAUC

とF値は，実際の値と予測の値を集計して計算される指標である。つまり，実際の値と予測の値のクロス表である。このクロス表は，混同行列（Confusion Matrix）と呼ばれる。以下に示す図2が，混同行列である。図の括弧内の値である0（陰性）と1（陽性）の2値を予測したとする。実際の値が1で予測の値も1であった場合，真陽性（True Positive：TP）と呼ばれる。また実際に0で予測も0であった場合，真陰性（True Negative：TN）と呼ばれる。これら予測結果は，正しく予測されたものである。誤って予測されたものは，実際の値は0であるが予測の値は1であった場合，偽陽性（False Positive：FP）と呼ばれる。また実際の値は1であるのに予測の値が0であった場合には，偽陰性（False Negative：FN）と呼ばれる。これらを使って，AUCでは対象のカテゴリの予測スコア（予測確率）を使って降順にソートし，真陽性（TP）が偽陽性（FP）よりも高くなるか測るものである。また，F値もこの実際の値と予測の値から計算され，適合率（Precision）と再現率（Recall）から計算される。Precisionは，TP/(TP＋FP）で計算され，予測値の1に対する真陽性の割合である。またRecallは，TP/(TP＋FN）で計算され，実際の値の1に対する真陽性の割合である。この適合率と再現率の2つの値から，F値は，2*Recall*Precision/(Recall＋Precision）で表される指標である。

図2　混同行列

		実際の値	
		1	0
予測値	1	TP	FP
	0	FN	TN

このように，正答率とモデル精度の指標であるAUC，F値を使って，学習用データと検証用データにおいて測定し，各モデル精度としてモデルの比較・検討を行う。

4　分　析　結　果

以下の表1は，予測器のアルゴリズムを使って大規模データについて予測判別の比較を行った結果である。左側のモデリング手法は今回使用したモデルで，CHAID，CART，C5.0とランダムフォレスト，XGBoost，ニューラルネットワークである。データ分割の割合は3つに分かれており，5：5，7：3と9：1の割合で，学習用データと検証用データに分割した。各テーブルの値は各手法の精度で，値はパーセンテージで表示されている。

まず5対5の分割，つまりデータの半分を学習用に使い，残りの半分を検証用データに分けた結果である。この結果，学習用データでの精度が高いモデルは，74.95％の精度で予測できているランダムフォレストのモデルである。しかし，検証用データの結果では，精度は

表 1　モデルの正答率（%）

モデリング手法	データ分割の割合 5：5 学習用	5：5 検証用	7：3 学習用	7：3 検証用	9：1 学習用	9：1 検証用
CHAID	71.56	70.38	71.89	70.37	72.07	71.01
CART	71.94	70.43	71.64	70.92	71.18	70.78
C 5.0	71.52	71.31	71.70	71.52	71.82	71.38
ランダムフォレスト	74.95	66.88	74.51	67.20	74.08	67.24
XGBoost	71.12	71.80	72.20	71.99	72.15	71.38
ニューラルネットワーク	71.65	70.95	72.15	71.31	72.18	71.22

66.88%となっており，検証用データの中では最も精度が低い値である。このことから精度は大幅に低下していることがわかる。これは，ランダムフォレストのモデルは学習用データで過学習してしまいオーバーフィットし，モデルの一般化がされていないことがわかる。学習用データと検証用データで高い精度をとるモデルは，XGBoost のモデルである。精度は，学習用データで71.12%，検証用データで71.80%と検証用データでも精度は上がっていることがわかる。その次に高い精度が C 5.0 である。学習用データで71.52%，検証用データで71.31%の精度である。C 5.0 の決定木のモデルでは検証用データの結果でも精度は大きくは落ちないことがわかる。CHAID と CART は，学習用で71.56%と71.94%，検証用では70.38%，70.43%の精度である。ニューラルネットワークでは，学習用が71.65%，検証用が70.95%である。学習用では，CHAID よりは高いが CART よりも低く，検証用データの結果では CHAID，CART よりも高い精度である。ニューラルネットワークのモデルでは，決定木よりも検証用データに対する精度が落ちていない。このことから決定木ほど過学習していない。以上の結果から，ランダムフォレストは学習用データでは精度が高いが，オーバーフィットする傾向があり，XGBoost の結果が学習用データ，検証用データとも精度が高い安定的な精度で予測できた結果である。また，他のデータ分割の結果，7 対 3，9 対 1 の結果でも，同様に XGBoost の結果が学習用，検証用とも安定した精度で，精度が大きく落ちていない結果となる。同じアンサンブル学習でも，ランダムフォレストのモデルは検証用データにおいて精度を大きく落としている。決定木では，C 5.0 の結果が予測精度は高い結果である。また，ニューラルネットワークのモデルは，学習用データの割合が高くなる 7 対 3，9 対 1 の精度を見ると，予測精度は高くなっており，学習データが増えることでのモデルの精度が高くなる傾向がある。検証用データに対して，精度は下がって71.22%であるが，XGBoost の71.38%に近い精度である。以上のことから，XGBoost の結果はモデルの安定性を図るのに良い結果であることがわかった。

以下の表 2 は，AUC，F 値のモデル評価の指標を示す。分析結果は，データを分割ごとに示し 5：5，7：3，9：1 の結果を表示する。モデリング手法と学習用データ，検証用データ

表２　モデル評価の指標

データの割合
＝5：5

モデリング手法	AUC 学習用	AUC 検証用	F 値 学習用	F 値 検証用
CHAID	0.784	0.754	0.672	0.655
CART	0.768	0.756	0.678	0.656
C 5.0	0.752	0.748	0.663	0.655
ランダムフォレスト	0.829	0.716	0.721	0.622
XGBoost	0.794	0.777	0.673	0.665
ニューラルネットワーク	0.782	0.767	0.675	0.663

データの割合
＝7：3

モデリング手法	AUC 学習用	AUC 検証用	F 値 学習用	F 値 検証用
CHAID	0.784	0.755	0.672	0.650
CART	0.767	0.762	0.673	0.661
C 5.0	0.750	0.747	0.659	0.652
ランダムフォレスト	0.825	0.720	0.713	0.623
XGBoost	0.790	0.781	0.674	0.666
ニューラルネットワーク	0.787	0.772	0.680	0.664

データの割合
＝9：1

モデリング手法	AUC 学習用	AUC 検証用	F 値 学習用	F 値 検証用
CHAID	0.785	0.758	0.677	0.661
CART	0.753	0.745	0.660	0.650
C 5.0	0.753	0.746	0.653	0.652
ランダムフォレスト	0.818	0.711	0.706	0.621
XGBoost	0.790	0.775	0.674	0.660
ニューラルネットワーク	0.786	0.769	0.678	0.663

で，AUC と F 値の値を表示する。まず AUC の値であるが，ランダムフォレストの値が高く，学習用データの中で，5：5 では0.829，7：3 では0.825，9：1 では0.818と他のモデルに比べ高い値である。しかし，検証用データでは 5：5 では0.716，7：3 では0.720，9：1 では0.711と他のモデルに比べて低い値である。これは上の精度でも述べた学習用データの過学習によるオーバーフィットした結果である。学習用データと検証用データとも安定した結果は，XGBoost の結果である。AUC の結果は，5：5 の分割で学習用0.794と検証用0.777，7：3 の分割で学習用0.790と検証用0.781，9：1 の分割で学習用0.790と検証用0.775である。また，ニューラルネットワークは，5：5 の分割で学習用0.782と検証用0.767，7：3 で学習用0.787と検証用0.772，9：1 で学習用0.786と検証用0.769である。決定木のモデルの CHAID では，5：5 の分割で学習用0.784と検証用0.754，7：3 では学習用0.784と検証用0.755，9：1 では学

習用0.785と検証用0.758である。AUC のモデル評価の指標結果では，XGBoost，ニューラルネットワーク，決定木の順となる。F 値の結果においても，ランダムフォレストの結果は学習データでは高い結果となる。学習用の結果は 5：5 の分割で0.721，7：3 では0.713，9：1 では0.706と高い値をとる。しかし，検証用データでは，5：5 の分割で0.622，7：3 の分割で0.623，9：1 の分割で0.621となり，低い結果である。F 値について XGBoost とニューラルネットワークの結果を 5：5 の分割で比較すると，XGBoost が学習用0.673，検証用0.665に対し，ニューラルネットワークは学習用0.675，検証用0.663の結果となった。XGBoost は，学習用では F 値が0.02低いものの，検証用では0.002高い結果であった。また，決定木では同じ 5：5 の分割で比較すると CART の結果は学習用で0.678と XGBoost とニューラルネットワークよりも高い，しかし検証用では0.656と XGBoost が0.009高い値であった。分割の割合が 7：3 では，XGBoost が学習用0.674，検証用が0.666に対して，ニューラルネットワークでは学習用が0.680，検証用では0.664である。CART では，学習用で0.673，検証用で0.661である。検証用の値で 3 つの手法を比較しても XGBoost がニューラルネットワークよりも0.002高く，CART よりも0.005高い。分割の割合が 9：1 でも，XGBoost の学習用が0.674，検証用の値が0.660である。ニューラルネットワークは，学習用が0.678，検証用では0.663である。CART は，学習用で0.660，検証用で0.650である。XGBoost の値は，検証用データにおいてニューラルネットワークの値よりも高くはない結果である。これは，学習用のデータが増えニューラルネットワークの精度が高くなっていることによるものと思われる。

5　ま　と　め

本論文では，“じゃらん”のデータを用いて各種の予測モデルの正答率やモデル評価の指標に関して比較し，検討を行った。予測モデルは，決定木として CHAID，CART，C5.0，アンサンブル学習としてランダムフォレストと XGBoost，そしてニューラルネットワークを用いた。予測精度，モデル評価の指標の比較を行った。データは10万件のデータをサンプリングし，学習用と検証用に分けて精度を比較した。主な分析結果は，以下の通りである。

1）伝統的な各種予測モデルの中で，決定木やニューラルネットワークに対して XGBoost が検証用データの正答率，モデル評価の指標に関して，良い結果をもたらした。

2）ランダムフォレストは，検証用のデータで大きく正答率やモデル評価の指標を落としてしまう結果であった。学習用データでは精度が高いことから，学習用データに対してオーバーフィットした結果であると考えられる。

3）C5.0 とニューラルネットワークを比較すると，学習用データと検証用データともに C5.0 が高い結果を得た。しかし，AUC や F 値に関しては学習用データと検証用データと

もに，ニューラルネットワークが高い値をもたらした。

4）CHAID や CART など伝統的な決定木の手法は，正答率に関して大きくは劣らないことがわかった。また，AUC や F 値などモデル評価の指標においても，これら手法が高い値をとることが明らかになった。

注

＊　本稿を執筆するにあたって，分析結果の利用を認めて頂いた株式会社リクルートライフスタイルに感謝申し上げます。本研究は JSPS 科研費 17H00983 の助成を受けたものです。

1）有効とは，予測モデルにおいて予測精度が高く，モデル精度の指標が高いモデルである。

2）このセクションは，麻生英樹・津田宏治・村田昇（2007），金（2017），馬場則夫・小島史男・小澤誠一（1994）を参考にした。

3）Hamori et al（2018）は，信用リスクの問題に対して，いくつかの機械学習の手法を応用し，比較検討を行っている。

4）使用するデータは，氏名や住所，電話番号等の個人に関する情報の個人情報や特定の個人を識別できる個人情報を含まない。

参　考　文　献

Ahmad, M. A., Mourshed, M. and Rezgui, Y.（2017）. “Trees vs Neurons: Comparison between Randomforest and ANN for high-resolution prediction of building energy consumption.” *Energy and Buildings* 147: 77-89.

Austin, P. C., Tu, J. V., Ho, J. E., and Levy, D., and Lee, D. S.（2013）. “Using methods from the data-mining and machine-learning literature for disease classification and prediction: a case study examining classification of heart failure subtypes.” *Journal of Clinical Epidemiology* Volume 66(4): 398-407.

Breiman, L.（1984）. *Classification and Regression Trees.* New York: Routledge.

Breiman, L.（2001）. “Random forests.” *Maching Learning* 45(1): 5-32.

Chen, T. and Guestrin, C.（2016）. “XGBoost: A Scalable Tree Boosting System.” *KDD ’16 Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*: 785-794.

Freund, Y. and Schapire, R. E.（1997）“A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting.” *Journal of computer and system sciences* 55: 119-139

Hamori, S., Kawai, M., Kume, T., Murakami, Y., and Watanabe, C.,（2018）“Ensemble learning or deep learning? Application to default risk analysis.” *Journal of Risk and Financial Management,* 11(1): 1-14.

IBM SPSS Modeler 18.1 Algorithms.（2017）.
ftp://public.dhe.ibm.com/software/analytics/spss/documentation/modeler/18.1/en/AlgorithmsGuide.pdf（最終アクセス：2019年2月28日）

Kass, G. V.（1980）. “An Exploratory Technique for Investigating Large Quantities of Categorical Data.” *Applied Statistics,* 29(2): 119-127.

Kurt, I., Ture, M. and Kurum, A. T. (2008). “Comparing performances of logistic regression, classification and regression tree, and neural networks for predicting coronary artery disease.” *Expert Systems with Applications* 34(1): 366-374.

McCulloch, W. S. and Pitts, W. (1943) “A logical calculus of the ideas immanent in nervous activity.” *The bulletin of mathematical biophysics* 5(4): 115-133.

Ozyirmidokuz, E. K., Uyar, K., and Ozyirmidokuz, M. H. (2015). “A Data Mining Based Approach to a Firm’s Marketing Channel.” *Procedia Economics and Finance* 27: 77-84.

Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning.* San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Razi, M. A. and Athappilly, K. (2005). “A comparative predictive analysis of neural networks (NNs), nonlinearregression and classification and regression tree (CART) models.” *Expert Systems with Applications* 29(1): 65-74.

Rosenblatt, F. (1958) “The perceptron: A probabilistic model for information storage and organization in the brain.” *Psychological Review* 65(6): 386-408

Rumelhart, D. E., Hinton, G. E. and Williams, R. J. (1986) “Learning representations by back-propagating errors.” *Nature* 323: 533-536

Tso, G. K. F., and Yau, K. K. W. (2007). “Predicting electricity energy consumption: A comparison of regression analysis, decision tree and neural networks.” *Energy* 32(9): 1761-1768.

Wendler, T. and Grottrup, S. (2016). *Data Mining with SPSS Modeler: Theory, Exercises and Solutions.* Switzerland: Springer International Publishing.

Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., Geoffrey J. M., Ng, A., Liu, B., Yu, P. S., Zhou, Z. H., Steinbach, M., Hand, D. J., and Steinberg, D. (2008) “Top 10 algorithms in data mining.” *Knowledge and Information Systems* 14(1): 1-37.

麻生英樹・津田宏治・村田昇（2007）『パターン認識と学習の統計学　新しい概念と手法』岩波書店。

金明哲（2017）『R によるデータサイエンスデータ解析の基礎から最新手法まで 第 2 版』森北出版。

馬場則夫・小島史男・小澤誠一（1994）『ニューラルネットの基礎と応用』共立出版株式会社。