成績・卒業を左右する要因について

荒　木　英　一*

1．はじめに

すでに旧聞に属する感もあるが，2016年6月3日毎日新聞に『大学卒業成績：1年終了時と関連』と題する記事が掲載されている。東日本のある理系大学での調査報告が紹介されており，1年終了時の成績と卒業時の成績が強い相関を持つことや，入試の形態・点数と入学後の成績には相関がないといったことが明らかにされたとある。

小論の目的は，いまあらためて，こうした問題を考察してみることである。

上の記事を発端にしていくつかの大学でも同様の分析結果が報告されたように記憶するが，それらはいずれも，本学とはランクも系統も異なる大学での調査報告であり，本学と類似する中堅文系私学における分析例を，筆者は寡聞にして知らない。本学においてもこうした分析を行ってみる意義はあろうかと考える。

また，本学経済学部が実施してきた新入生アンケート調査（E-folio）の回答結果が相当に蓄積されており，これに全学事務システムやWebから取得したデータを加えることで，限定的ながら，こうした分析に耐えうるデータフレームの構築が可能となったことも，小論の動機となっている。

＊ 事務システムからのデータ収集にあたって中野瑞彦教授にご尽力いただいた。同時に，事務各所管からもデータ提供の快諾をいただいた。また，桃山学院大学総合研究所共同研究プロジェクト（20共275）「経済学部独自アンケート（E-folio）の深化に向けて」のメンバーからも有益なコメントをいただいた。記して，深謝する。なお，言うまでもなく，ありうべき過誤は筆者ひとりの責に帰す。

キーワード：経済学部生，成績，卒業，決定木，ランダムフォレスト

小論では，以下の 3 点について考察をしたい。

- 1 回生終了時の成績を左右する要因は何か

  大学での成績に，出身高校の偏差値や評定平均値，さらに入試区分がもたらす影響はいかほどか。また，新入生アンケートに（部分的にでも）あらわれるはずの学生ひとりひとりの差異がもたらす影響はいかほどか。

- 4 回生終了時の成績を左右する要因は何か

  「1 回生時の成績が 4 回生時の成績を決める」という命題は本学でも成立するか。4 年間の大学生活を経て，出身高校偏差値や評定平均値，入試区分等による差異は消滅するか。

- 4 年で卒業できるか否かを左右する要因は何か

  4 年間での卒業の成否を，1 回生時の実績のみから予測することは可能だろうか。

分析は，経済学部新入生アンケート（E-folio，2012〜19 年）の回答に，全学事務システムから抽出された教務入試関連データをマージ（併合）し，さらにWebから取得した若干のデータを加えて，行われた[1)]。分析に用いた手法は，線形回帰（OLS），回帰木（Random Forest），分類木である。

2．1 回生終了時の成績を左右する要因

まず，1 回生終了時点でのGPAに着目して，これの決定要因を考察しよう。

1 回生終了時点でのGPAを目的変数として，今回準備したデータフレーム上で利用可能な説明変数を網羅してみる。このモデルを，解釈がしやすい線形回帰（OLS）によって推計した結果が，図 1 である[2)]。

表の上から順に，（0.1% もしくは 1% 水準で）有意な説明変数を列挙す

1）Webからの全国高校偏差値データの取得や，アンケート回答（自由記述欄）のテキスト処理などで，いくぶん煩雑な前処理を経ている。

2）対象は，2012 年度から 2018 年度までの本学経済学部入学生（2019 年度入学生はデータ収集時点でまだ 1 回生秋学期に在籍中）である。

	Estimate	Std.Error	t value	Pr(>|t|)	
(定数項)	-1.66777	0.20056	-8.316	< 2e-16	***
出身校偏差値	0.02114	0.00253	8.344	< 2e-16	***
評定平均値	0.45333	0.02821	16.072	< 2e-16	***
入門演習の成績 (0,1,2,3,4)	0.22789	0.01508	15.118	< 2e-16	***
AO入試 (1/0 Boolean)	-0.24577	0.05765	-4.263	2.13e-05	***
公募制入試 (1/0 Boolean)	-0.04880	0.03749	-1.302	0.19323	
指定校入試 (1/0 Boolean)	-0.14143	0.04054	-3.489	0.00050	***
スポーツ推薦入試 (1/0 Boolean)	-0.34817	0.08073	-4.313	1.71e-05	***
1: 自宅 ／ 0: 下宿	-0.06040	0.05438	-1.111	0.26690	
通学時間 (1.0/1.5/2.0)	0.17526	0.05174	3.387	0.00072	***
学生生活満足度 (1,2,3,4,5,6,7,8,9,10)	0.00813	0.00740	1.099	0.27180	
大学生活に不安あり (1/0 Boolean)	-0.09485	0.04034	-2.351	0.01881	*
どんな不安か (回答文の字数)	0.00512	0.00251	2.044	0.04115	*
単位／授業／試験に不安あり (1/0 Boolean)	0.12743	0.04072	3.129	0.00178	**
経済学部で学びたいことがある (1/0 Boolean)	-0.06129	0.04316	-1.420	0.15583	
何を学びたいか (回答文の字数)	0.00729	0.00310	2.352	0.01881	*
大学でチャレンジしたいことがある (1/0 Boolean)	0.08184	0.03869	2.116	0.03453	*
何にチャレンジしたいか (回答文の字数)	0.00382	0.00246	1.554	0.12037	
春学期に興味を持った科目がある (1/0 Boolean)	0.01692	0.04831	0.350	0.72622	
経済学関連の科目に興味をもった (1/0 Boolean)	0.09255	0.03392	2.728	0.00643	**
いまアルバイトをしている (1/0 Boolean)	-0.07435	0.02919	-2.547	0.01094	*
すでに何らかの資格を取得している (1/0 Boolean)	-0.18265	0.15078	-1.211	0.22592	
大学でのサークル部活動に参加 (1/0 Boolean)	-0.07558	0.02983	-2.534	0.01137	*
CBCCコースに所属 (1/0 Boolean)	1.05310	0.11935	8.823	< 2e-16	***

*** 0.1%有意, ** 1%有意, * 5%有意 .10%有意

Residual standard error: 0.5758 on **1686** degrees of freedom
(1113 observations deleted due to missingness)
Multiple R-squared: **0.3267**, Adjusted R-squared: **0.3175**
F-statistic: 35.56 on 23 and 1686 DF, p-value: < 2.2e-16

図1　1回生終了時のGPA(OLS)

ると，まず，出身校偏差値，評定平均値，さらに「入門演習」の成績[3]，この3つはすべて正の係数推定値となる。

続く4つは入試区分をあらわすダミー変数で，(公募制をのぞく) 推薦入試合格者について負の効果が確認できる[4]。

これより下の説明変数は，2012年から19年までの春学期（5月初旬）に実施された「経済学部新入生アンケート（E-folio)」の回答結果である。通

3) 1回生春学期配当のいわば大学生活導入科目であり，ここではその成績 S，A，B，C，D をそれぞれ4，3，2，1，0としている。

4) たとえばスポーツ推薦合格者はGPAが0.35ほど低く，AO合格者は0.25ほど低くなる。

学時間（1時間未満を1，1〜2時間を1.5，2時間以上を2としている）が正の係数，また「大学生活を送るうえでなんらかの不安があるか」という自由記述の質問に対して「単位」「授業」「試験」といったキーワードを含む回答をした場合に正の係数，また「春学期に興味を持った科目があるか」との自由記述の質問に対して経済学関連の科目を回答した場合にも正の係数が確認できる。これら以外にも5%水準で有意な説明変数があるが，係数の符号はおおむね妥当なものと解釈できるだろう。

また，CBCCコース（中国ビジネスキャリアコース，2015年度末に発展的解消）の学生たちのGPAは，他の経済学部生全体より平均で1.05ポイント高かったことがわかる。

ただし，この回帰式全体の決定係数は0.3175と低い。

この線形回帰の結果を，別の手法からも確認しておこう。

図2は，上と同じモデルを，回帰木（ランダム・フォレスト）で処理した結果である。

回帰木とは目的変数が連続値をとる場合の決定木分析をいう。また，与えられた標本から（ブーツトラップにより）ランダムにサンプルレコードを抽出しながら決定木分析を反復する処理をバギング（Bootstrap　AGGregation ING）といい，ランダム・フォレストではさらに説明変数（特徴量）の組み合わせもランダムに変更しながら決定木分析を反復する[5)]。ある説明変数を（ランダムに選ばれた他の説明変数との組み合わせのもとで）加えたり外したりしながら，その変数が目的変数の予測精度の向上にどれくらい貢献しているかを計測したものが，図2に示す「重要度（%IncMSE）」である[6)]。

上の線形回帰と同様に，評定平均値，出身校偏差値，入門演習の成績，

5）小論の推計では，いずれも，毎回ランダムに選抜される説明変数の個数を7個，反復処理の回数を500回（デフォルト値）とした。

6）より詳しくは，重要度指標の%IncMSE（Percentage Increase in Mean Square Error）とは，OOB（Out of Bag，バギングの際に選ばれなかったレコード）を用いて推計モデルの交差検証をする際に，当該説明変数の値を意図的に変更（シャッフル）して平均二乗誤差の増加分（悪化分，パーセント値）を算定したものである。

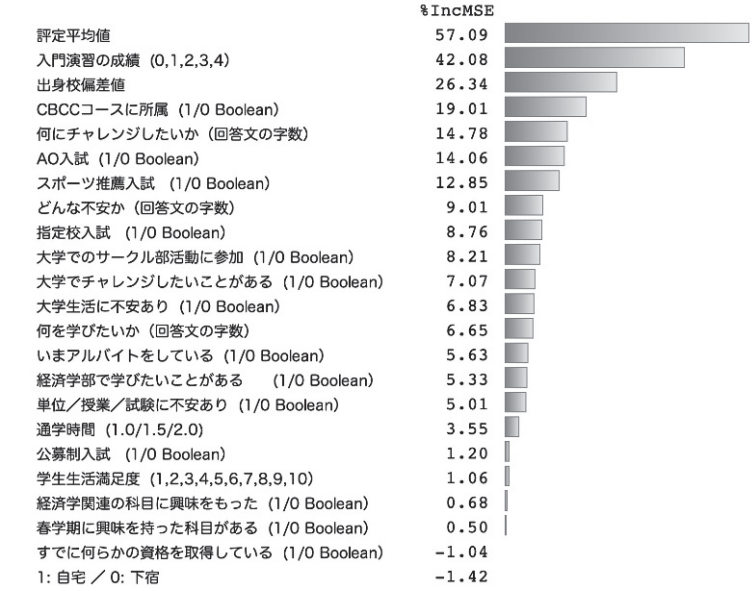

図２　１回生終了時のGPA（回帰木Random Forest）

CBCCコース所属，そして推薦入試合格を示す変数が，重要度の上位にランクされており，線形回帰からの結果が追認できる。

また，この推計モデルの説明力は 27. 37% で，線形回帰と同様に 30% 弱にとどまる。

１回生終了時のGPAを決定する要因として，評定平均値，出身校偏差値，入試区分は重要である[7)]。

ただし，その影響度はたかだか 30% にとどまる。線形回帰，回帰木（ラ

7）この結論は妥当にも思えるが，たとえば「入試の形態・点数と一回生時成績は無関係」という，他所での調査結果とはいくぶん異なる。

ンダム・フォレスト）ともに，説明力の高いモデルは得られていない。これは，なんらかの重要な説明変数の欠落を含意するものと考え得る。

3．4 回生終了時の成績を左右する要因

次に，4 回生終了時点でのGPAに着目して，これを決定する要因を考察したい。

4 回生終了時点でのGPAを目的変数として，前節と同じ説明変数を網羅し，さらに 1 回生終了時点でのGPAを説明変数に加えてみる。このモデルを線形回帰（OLS）によって推計した結果が，図 3 である[8)]。

	Estimate	Std.Error	t value	Pr(>|t|)	
（定数項）	0.27000	0.12810	2.108	0.03526	*
GPA（一回生終了時）	0.65538	0.01659	39.507	< 2e-16	***
出身校偏差値	0.00495	0.00166	2.981	0.00294	**
評定平均値	0.08771	0.01884	4.655	3.60e-06	***
AO入試（1/0 Boolean）	0.04599	0.03806	1.209	0.22708	
公募制入試（1/0 Boolean）	0.04275	0.02309	1.851	0.06437	.
指定校入試（1/0 Boolean）	0.02366	0.02839	0.834	0.40464	
スポーツ推薦入試（1/0 Boolean）	-0.05905	0.05638	-1.047	0.29513	
1: 自宅 ／ 0: 下宿	0.03541	0.03822	0.927	0.35432	
通学時間（1.0/1.5/2.0）	-0.06212	0.03304	-1.880	0.06031	.
学生生活満足度（1,2,3,4,5,6,7,8,9,10）	0.00476	0.00470	1.012	0.31170	
大学生活に不安あり（1/0 Boolean）	0.03394	0.02486	1.365	0.17244	
どんな不安か（回答文の字数）	-0.00260	0.00138	-1.882	0.06002	.
単位／授業／試験に不安あり（1/0 Boolean）	0.01923	0.02521	0.763	0.44590	
経済学部で学びたいことがある　（1/0 Boolean）	-0.01284	0.02701	-0.475	0.63465	
何を学びたいか（回答文の字数）	0.00032	0.00108	0.294	0.76918	
大学でチャレンジしたいことがある（1/0 Boolean）	0.00882	0.02791	0.316	0.75218	
何にチャレンジしたいか（回答文の字数）	0.00262	0.00133	1.969	0.04921	*
春学期に興味を持った科目がある（1/0 Boolean）	-0.02130	0.03133	-0.680	0.49672	
経済学関連の科目に興味をもった（1/0 Boolean）	0.02600	0.02303	1.129	0.25904	
大学でのサークル部活動に参加（1/0 Boolean）	-0.09281	0.01965	-4.723	2.60e-06	***
CBCCコースに所属（1/0 Boolean）	0.34196	0.04736	7.221	9.27e-13	***

*** 0.1%有意，** 1%有意，* 5%有意　.10%有意

Residual standard error: 0.3183 on **1176** degrees of freedom
(1625 observations deleted due to missingness)
Multiple R-squared: **0.6766**,　Adjusted R-squared: **0.6709**
F-statistic: 117.2 on 21 and 1176 DF,　p-value: < 2.2e-16

図 3　4 回生終了時のGPA（OLS）

8）対象は，2012 年度から 2015 年度までの本学経済学部入学生（2016 年度以降の入学生はデータ収集時点でまだ 4 回生秋学期に在籍中）である。

まず，1回生終了時GPAが0.1% 水準で有意（正の回帰係数）となる。また，（4年前に卒業した出身高校の）偏差値と評定平均値の影響は残存しており，いずれも（1% もしくは0.1% 水準で）有意となっている。

入試区分に関する説明変数は有意ではなくなっており，4年の在学期間を経て入試区分による差異は消滅すると解釈したい。

しかし，4年前（1回生時）の5月に行われたアンケート調査の結果は，「大学でのサークル部活動に参加」が1% 水準で有意（負の回帰係数），「何にチャレンジしたいか」という質問に（新入生時にすでに具体的に）回答していた場合には5% 水準で有意（正の係数）である。

さらに，CBCCコースへの所属はここでも（0.1% 水準で）有意となり，CBCCコース所属学生のGPAが0.34ほど高くなっている[9)]。

回帰の決定係数は0.67と高いが，すぐあとにも見るように，1回生終了時GPAの貢献がとりわけ大きい。これは，「1回生時の成績がその後の成績を大きく左右する」という従来からの指摘に合致するものでもある。これに比べると，限界的な影響は大きくないが，出身高校偏差値と評定平均値の影響も有意に残存する。

前節と同様に，この線形回帰から得られる解釈を，回帰木（ランダム・フォレスト）による推計でも確認してみよう。

図4は，同じモデルを回帰木（ランダム・フォレスト）で処理した結果である[10)]。

図4より，4回生終了時のGPAを予測するうえで，1回生終了時GPAの重要度がきわだって高いことがわかる。高校時代の評定平均値の影響も，いぜ

9）CBCCコースに在籍した者のGPAは，1回生終了時にも4回生終了時においても，有意に高い。もちろん，出身校偏差値や評定平均値をはじめ他の要因の影響をコントロールした上でもなお有意に高いのであるが，参考までに，CBCCコースに在籍した学生の出身校偏差値の平均は43.87であり，それ以外の学部生の平均45.11を下回る。ちなみに，評定平均値の平均は3.61で，それ以外の学部生平均3.56より少し高い。

10）説明変数として，「1回生時に低単位取得者であったかどうか」というダミー変数を加えている。「低単位取得」の定義としては（便宜的に）全体の単位取得数の25% 点（＝10単位）未満とした。

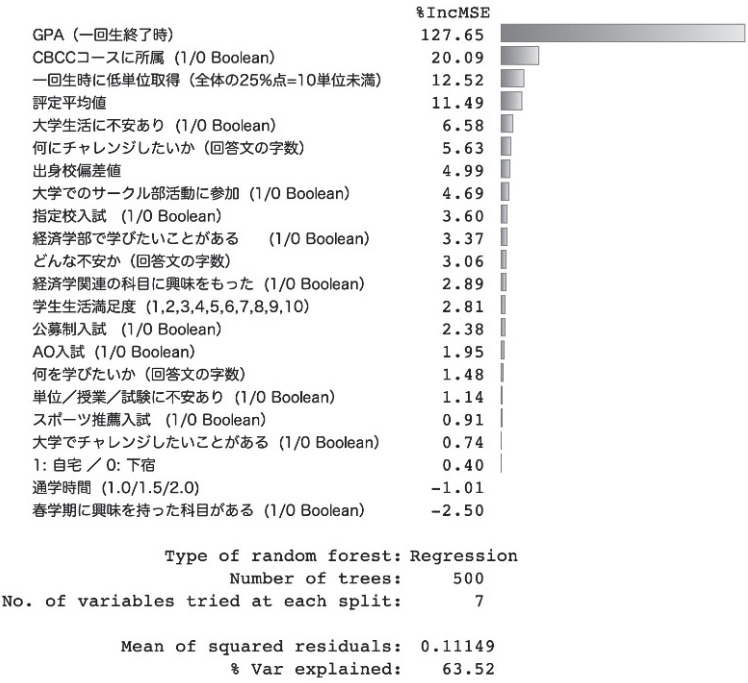

図 4　4 回生終了時のGPA（回帰木Random Forest）

んとして残存する。また，E-folio回答項目のうち，上の線形回帰で有意と判定した説明変数はやはり上位にランクされている[11]。

また，この推計モデルの説明力は 63.52% で，線形回帰と同様に 60% を超える。

4．4 年間で卒業できるか否かを左右する要因

本学部において，4 年間で卒業できた学生の比率は，2012 年度入学生から，72.37%→73.27%→76.10%→72.54% と推移している（データ収集時

11）E-folio回答項目のうち，「通学時間」はOLSでは 10% 有意で負の影響を持つが回帰木では重要でない変数（落としたほうが予測誤差は改善）となり，逆に「大学生活に不安あり（Yes／No選択）」は回帰木では重要となるがOLSでは有意でない。

点では2015年度入学生までのデータが利用可能)。

この節では，4年後の卒業の成否を1回生時点（春学期終了時点）の指標から予測可能かどうかについて，検討したい。

同種のものとして1回生春学期中の指標をもとに数年先までの中退（退学・除籍）予備軍を予測する分析が他所で行われているが，本節では，目的変数として，4年で卒業できた場合を1とし，できなかった場合を0とするブール（ダミー）変数をとることにしたい[12)]。

説明変数としては，第2節で一回生終了時のGPAを説明するために用いたものをすべて投入し，さらに，1回生春学期終了時のGPA（変数名はS1G）もしくは取得単位数（変数名はS1U）を説明変数に加えてみる。

まず，木構造がはっきり見通せて解釈もしやすい，シンプルな決定木（分類木）による分析を試みよう[13)]。

図5が，誤分類率が最小になる木のなかで，最もシンプルな分類木である。登場する変数は，S1U（1回生春学期終了時の取得単位数）と評定平均値，そしてWSD（E-folio新入生アンケートの自由記述型の質問「経済学部で学びたいことがあるか」にポジティブに回答したか否か）の3変数のみである。

右上のルートノード1から出発して，サンプル総数1181名が5個のグループ（終端ノード）に分類されていく。まず，S1U≧9.5ならば右下の終端ノード（Node 9）におさまる（954名がここに分類され，4年で卒業できる確率は83.96%）。S1U<9.5なら左のノード2へ進み，さらにS1U<4.5ならば左下の終端ノード（Node 3）に分類される（52名がここに分類され，4年で卒業できる確率は15.38%）・・・と分類されていく。

12）つまり，退学・除籍に加えて留年した場合も0とする。

13）たとえば須田[2018]等では，まずランダム・フォレストで重要な説明変数をきり出し，それらだけを使って決定木をあらためて求めるという手法がとられているが，本節の分析では，第一段階のランダム・フォレストをスキップしても充分にシンプルな木構造が得られた。シンプルな木構造を求めることは重要であると思われる。なぜなら，こうした分析結果は，実際になんらかの介入を行う際の意思決定資料となりうるからである。

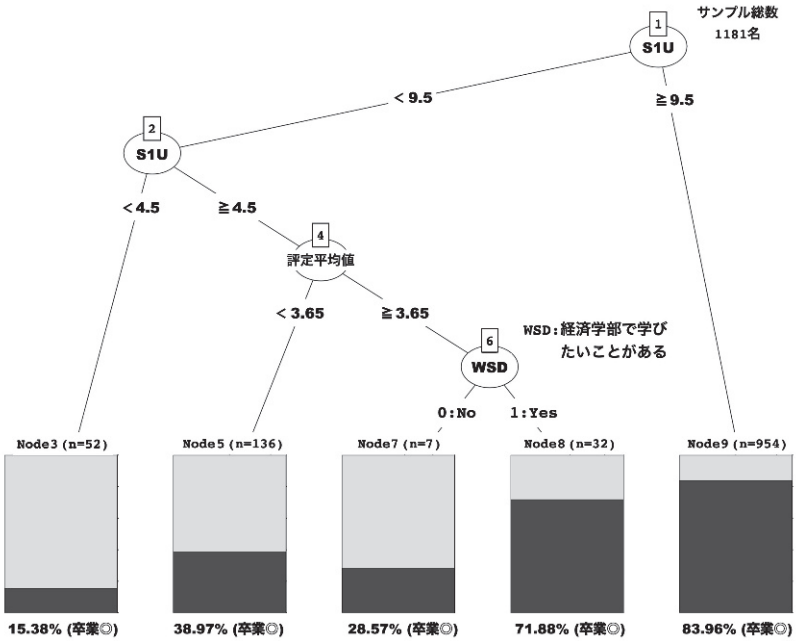

図 5　◎:4 年間で卒業できる ／ ×:できない
（S 1 U:1 回生春学期終了時の取得単位数）

図 6 は，この分類木の誤分類率を表にまとめたものである。1181 名のサンプル中 956 名が正しく分類され，全体の誤分類率は 19.05% にとどまっている。また，実際には 4 年で卒業できた者を「卒業できなかった」と誤分類したケースは 7.1% にとどまる。

	実測値	◎に分類	×に分類	誤分類率
1: 4年で卒業◎	887	824	63	7.10%
0: 4年で卒業×	294	162	132	55.10%

サンプル数	1181	うち正しく分類された者	956
		全体の正答率	80.95%
		全体の誤分類率	19.05%

図 6　誤分類率(分類木)

しかし，「実際には卒業できなかった者」を「卒業できた」と誤分類したケースが 55.1% にのぼっている（つまり正答率は 44.9%）。同じモデルをランダム・フォレスト等で処理した場合にも，この正答率の改善は見られな

かった。

他所での類似研究における正答率を見ると，たとえば近藤・畠中[2016]の場合，1回生春学期第5週終了時点の指標を用いて，それらを三層のニューラルネットワーク（RBFN）に投入すれば，3回生春までにドロップアウト（中退）する学生のおよそ40% を予測できるという。また，1回生春学期終了時点の指標を用いるならば，多くの機械学習手法によってドロップアウトの50〜60% を予測できるという。

これらに比して本節モデルの予測精度はいくぶん低く，改良の余地が残されている。手法の見直しもさることながら，有効な説明変数の探索および取得（入手可能性）がより重要であろうと考える[14]。

5．結びにかえて

小論では，本学経済学部生の成績と卒業に関して，以下の分析結果を得た。

(1) 1回生終了時のGPAを左右する要因として，出身高校偏差値と評定平均値は重要である。どの入試区分で合格したか（推薦入試か否か）も重要な要因である。しかし，それらの影響はたかだか30% にとどまる。また，E-folio調査のいくつかの項目も有意に整合的な影響を示している。

(2) 4回生終了時のGPAを左右する要因として，1回生終了時のGPAがきわめて重要である。出身高校偏差値と高校時代の評定平均値の影響も，いぜんとして残存する。4年前に行われた新入生アンケート（E-folio）調査項目のいくつかも有意に影響している。

(3) 1回生春学期終了時点の指標とE-folio調査項目を組み合わせて，卒業の成否を予測できる可能性がある。そのためには，有効な説明変数の探索および取得が重要と思われる。

14) ちなみに，近藤・畠中[2016]が用いている指標は，性別・入試区分・入学前学習提出度・オリエンテーション出席度・1〜5週目出席率・6〜10週目出席率・11〜15週目出席率・春学期GPAである。これらのうち，性別・入試区分・春学期GPA以外の変数は，小論の今回の分析では利用できていない。

残された課題の検討と，より包括的な分析については，別稿を期したい。

参考文献

[1] 近藤伸彦，畠中利治「学士課程における大規模データに基づく学修状態のモデル化」2016，教育システム情報学会誌 33 巻 2 号　pp. 94−103

[2] 白鳥成彦，大石哲也，田尻慎太郎，森雅生，室田真男「中退確率の遷移を用いた中退学生の類型化」2020，日本教育工学会論文誌 44 巻 1 号　pp. 11−22

[3] 須田茂夫「機械学習による都道府県別医療費の分析」2018，社会保障研究 3 巻 3 号pp. 403−415

[4] S.Athey, J.Tibshirani and S.Wager, “Generalized Random Forests”, *The Annal of Statistics,* 2019, Vol.47, No.2, pp. 1148−1178

[5] K.Ito, T.Ida and M.Tanaka, “Moral Suasion and Economic Incentives: Filed Experimental Evidence from Energy Demand”, *American Economic Journal: Economic Policy* 2018, 10(1), pp.240−267

[6] 荒木英一，福本幹生「経済学部E-folio調査にもとづく『やる気』の考察」2015，桃山学院大学経済経営論集 57 巻 1 号　pp. 19−36

（あらき・えいいち／経済学部教授／2020 年 9 月 25 日受理）

Factors Affecting the Academic Performance and Graduation of University Students

ARAKI Eiichi

This article investigated what factors affect student's academic performance and successful graduation within 4 years at the faculty of economics of a medium-sized private university. Our faculty has been conducting a relatively large-scale questionnaire to freshmen students since 2012. Based on each student's answers to the questionnaire combined with their official academic records, we made an analysis using random-forest regression/classification techniques.

We have got the following propositions.

1. GPA in freshman year is affected by the academic ranking of the high school each student graduated from, GPA in their high school days, and whether their admission is based on recommendation or not. In addition, student's degree of motivation, which can be drawn from their answers to the questionnaire, is also significant. However, these factors can explain at most only 30% of the variation in GPA in freshman year.
2. GPA in senior year deeply depends on GPA in freshman year. Both are in a strong positive correlation (ρ=0.8). In addition, GPA in student's high school days still remains significant. These factors explain more than 60% of the variation in GPA in senior year.
3. Prediction about which students will drop out becomes possible at the end of the first semester in freshman year. Although the accuracy of our simple model is around 45%, we hope it can be improved by incorporating more explanatory variables into the model.