O-012

入学前教育データを用いた大学入学後の成績予測のための複数の機械学習モデルの比較検討

Comparison of Several Machine Learning Models for forecasting Post-enrollment Grades Using Pre-admission Education Data

荒澤 孔明† 松川 瞬† 杉尾 信行† 高原 まどか†† 服部 峻†††
Komei Arasawa† Shun Matsukawa† Nobuyuki Sugio† Madoka Takahara†† Shun Hattori†††

1. はじめに

日本の大学では，中途退学者の増加が深刻化している．学生の退学理由には，成績不振が大きな要因であると主張する研究者らも多くおり，早い段階で学生の成績を予測し，教員らが早期にサポートできる仕組みの確立が求められている[1]．本稿では，入学前教育の学修記録や高校在籍時の成績，また入試情報などを説明変数とした機械学習手法に基づき，大学の各合格者の 1 年後の GPA の順位区分が下位$1/N$（$N = 2,3,4$）になるか否かを予測する複数のモデルの構築を行い，その予測性能の比較を行う．

2. 提案手法

2.1 説明変数

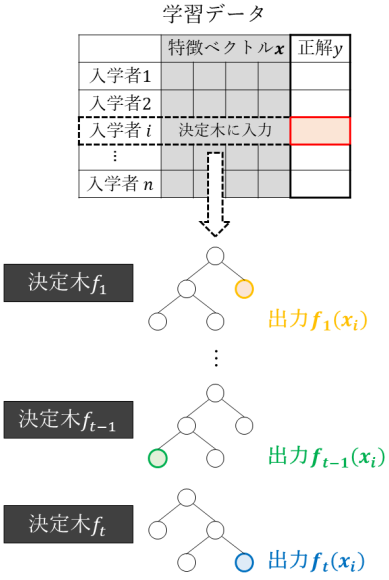
大学合格者の入学後の成績予測を行うために検討した説明変数は表 2 の 16 種類である．このうち，入学前教育データとは，大学入試の総合型選抜と推薦型選抜における合格者に課したオンラインドリルの約 3 ヵ月分の記録である．指定した教材は，難易度の基づきベーシックコースとステップアップコースに分かれている．各コースには，5 科目（英・国・社・数・理）用意されており，さらに各科目の中には 6 単元用意されている．受講者は計 30 単元分のドリルと小テストを実施する．なお，各単元のドリルと小テストは何回でも実施できる．ここでは，単元ごとの小テスト平均点，ドリル実施時間，ドリル実施回数が，2 コースぞれぞれで記録される．

2.2 予測モデル

本稿では，機械分類タスクにおいて高性能であり，目的変数と関係が強い説明変数の把握が容易である勾配ブースティング決定木を用いる．これは複数の決定木$\mathcal{F}$を直列に学習するモデルであり，前の学習器の結果をその次の学習器に反映させる特徴を持つ．ここで，i番目の入学者の 16 次元の特徴ベクトルを$\boldsymbol{x_i}$，それに対する目的変数を$y_i = \{0,1\}$とする．この時，その入学者の GPA 順位区分が下位$1/N$（$y_i = 1$）となる確率p_iは，各学習器$f_k(k = 1,2,\cdots,K)$の出力$f_k(x_i)$の総和をシグモイド関数で変換した値である．

$$p_i = \frac{1}{1 + exp(-\hat{y}_i)}$$

$$where, \hat{y}_i = \sum_{1 \le k \le K} f_k(x_i), f_k \in \mathcal{F}$$



決定木f_tは以下の損失関数$\boldsymbol{L^t}$が最小になるように作成

$$\boldsymbol{L^t} = \sum_{i=1}^{n} \boldsymbol{l}(\underbrace{f_1(x_i) + f_2(x_i) + \cdots + f_{t-1}(x_i)}_{\text{1つ前の決定木までの出力の合計}} + \underbrace{f_t(x_i)}_{\text{＋自身の木の出力}}, \underbrace{y_i}_{\text{正解ラベル}})$$

ただし，出力$f(x)$はT個の葉を持つ決定木に説明変数$\boldsymbol{x_i}$が入力された時，その決定木の分岐を下り，最終的に辿り着くq番目の葉の重み$w_{q(x)}$の事である．

$$\mathcal{F} = f(x) = w_{q(x)} \ (q: \mathbb{R}^{16} \to T, w \in \mathbb{R}^T)$$

続いて，各決定木の最適化について述べる．学習器f_tを最適化する際には，その 1 つ前までの学習器$f_k(k = 1,2,\cdots,t-1)$は既に最適化済みである．それを踏まえた上で，学習器f_tは，その 1 つ前までの学習器の出力の和（定数）に自信の出力（変数）を加えた値を算出し，この値と正解ラベルy_iとの誤差L^tを最小化するように最適化される．

$$L^t = \sum_{1 \le i \le n} l(\hat{y}_i^{(t-1)} + f_t(x_i),\ y_i)$$

ただし，l は損失関数で，交差エントロピーを示す．

$$l\left(\hat{y}_i^{(t-1)} + f_t(x_i),\ y_i\right) = l\left(\hat{y}_i^{t},\ y_i\right)$$
$$= y_i \log p_i + (i - y_i)\log(i - p_i)$$

3. 実験環境

実験では，北海道科学大学の各新入生の 1 年後の GPA の順位区分が下位 1/2 になるか否かの 2 値分類モデル，下位 1/3 になるか否かの 2 値分類モデル，下位 1/4 になるか否かの 2 値分類モデルの構築を行い，予測性能を評価する．

† 北海道科学大学 Hokkaido University of Science
†† 龍谷大学 Ryukoku University
††† 滋賀県立大学 The University of Shiga Prefecture

属性データ	1	性別	{“男”, “女”}
出身高校データ	2	出身高校の課程種別	{“普通科”, “工業科”, …など}
	3	出身高校の進路区分	※具体的な値は非公表とする
	4	高校在籍時の平均評定値	0 以上 4 以下
入試データ	5	入試区分	{“総合型選抜”, “推薦型選抜”, …など}
入学前教育データ	6	基礎学力テスト①の解答時間（分）	0 以上
	7	基礎学力テスト①の合計得点	0 以上 125 以下（25 点×5 科目）
	8	ベーシックコースにおける各単元の小テストの平均点の合計	0 以上 3000 以下（100 点×30 単元）
	9	ベーシックコースにおけるドリル総実施時間（分）	0 以上
	10	ベーシックコースにおけるドリル総実施回数	0 以上
	11	ステップアップコースにおける各単元の小テストの平均点の合計	0 以上 3000 以下（100 点×30 単元）
	12	ステップアップコースにおけるドリル総実施時間（分）	0 以上
	13	ステップアップコースにおけるドリル総実施回数	0 以上
	14	基礎学力テスト②の解答時間（分）	0 以上
	15	基礎学力テスト②の合計得点	0 以上 125 以下（25 点×5 科目）
	16	基礎学力テスト①と②の得点差	0 以上 125 以下

3.1.1 データセット

本稿では，北海道科学大学の学部ごとに，2021 年度の新入生データセット，2022 年度の新入生データセット，2023 年度の新入生データセットを用意した．なお，2021 年度のデータセットを学習データにし，2022 年度のデータセットをテストデータにする実験①と，2022 年度のデータセットを学習データにし，2023 年度のデータセットをテストデータにする実験②を行っている．データセットには，総合型選抜また推薦型選抜の合格者のみが含まれている．一般選抜の合格者を含めなかった理由は，入学前教育を行っていないためである．また，各年度の新入生の 1 年後の成績の分布は表 2 であり，順位は一般選抜入学者も含む自身の所属学科における順位である．

3.1.2 機械学習のフレームワーク

本稿では，機械学習のフレームワークとして，勾配ブースティング決定木の 3 手法（XGBoost[2] / LightGBM[3] / CatBoost[4]）と決定木ベースではない 2 手法（SVM，ニューラルネットワーク）を採用した．パラメータ調整は先行研究[5]のとおりである．

表 2 各年度における新入生の 1 年後の成績区分

		下位 1/2	以外	下位 1/3	以外	下位 1/4	以外
2021	工学	134	98	83	149	66	166
	薬学	32	34	25	41	19	47
	保健	88	82	60	110	46	124
	未来	47	38	32	53	25	60
	全学	301	249	200	350	156	394
2022	工学	258	171	184	245	142	287
	薬学	51	53	32	72	25	79
	保健	89	84	55	118	46	127
	未来	47	41	29	59	24	64
	全学	316	261	208	369	166	411
2023	工学	124	77	84	117	60	141
	薬学	66	53	37	82	30	89
	保健	83	75	56	102	36	122
	未来	34	42	21	55	14	62
	全学	307	244	198	353	140	411

4. 実験結果

表 4 は，①と②のデータセットを用いた時の機械学習手法の F 値である．ここから，下位 1/2 になる学生か否かを予測するモデルでは 0.7 前後，下位 1/3 の予測モデルでは 0.6 前後，下位 1/4 の予測モデルでは 0.5 前後の F 値が得られた．また，サポートベクターマシンやニューラルネットワークといった他のアルゴリズムと比較し，データの特性や構造が複雑な場合でも良い性能を示す事が多いとされている勾配ブースティング決定木の方が，やはり予測性能に優れており，これは直感と一致する結果であった．

また，表 5 と表 6 は，①と②のデータセットを用いた時の機械学習手法の適合率と再現率である． ここから，適合率の方が再現率よりも優れている傾向が窺える．これについて我々は，適合率が重視された予測モデルは，業務要件に適していると考えている．理由は 2 つある．

第 1 に，偽陰性が起きた時の問題が小さいためである．この結果のように予測モデルの再現率が低く，学業に問題がある入学者を見逃してまった場合でも，入学後に成績が芳しくないと疑われた場合，本研究の目的にはそぐわないが，その都度その学生を教員らがフォローアップする事が従来通りであり，指導漏れのリスクは少ないという事である．さらに，既存の入学後のデータを用いる成績予測システムでも補う事ができる．

第 2 に，偽陽性が起きた時の問題の方が大きいためである．大学の初年度は，新しい環境に適応したり，新しい行動や外見に挑戦したりする時期であり，入学前の学力や学業に対する意識が，1 年かけ変化する事も十分ありえる．さらに，大学の成績は，講義の受講態度，正課外活動の多さ，友人との付き合い方などといった，入学後の様々な要因によって左右される．すなわち，入学後の成績を入学前に予測しようとするタスクの難易度は非常に高い．そのため，偽陽性のリスクを追ってまで，1 年後に成績不振となる学生らを過度に検出してしまうと，問題ないはずである学生らまで過度な指導の対象となり，結果として，教員の負担だけでなく，学生の負担も増やしかねないという問題が生じてしまうという事である．

表 3 複数の機械学習手法における F 値の比較（ボールドはその行の最高値）

		勾配ブースティング決定木						決定木ベースではない手法			
		XGBoost		LightGBM		CatBoost		SVM		NN	
		①	②	①	②	①	②	②	②	②	②
下位 1/2 予測モデル	工学	0.707	0.683	0.711	0.703	0.709	0.704	0.495	0.686	0.495	**0.766**
	薬学	0.568	0.550	0.587	0.614	0.578	0.585	0.571	0.527	0.505	**0.717**
	保健	0.663	**0.692**	0.644	0.647	0.633	0.611	0.556	0.585	0.144	0.000
	未来	**0.777**	0.558	0.682	0.578	0.65	0.557	0.568	0.706	0.526	0.467
	全学	**0.702**	0.681	0.688	0.681	0.687	0.676	0.615	0.670	0.596	0.675
下位 1/3 予測モデル	工学	0.605	0.588	**0.662**	0.564	0.615	0.543	0.531	0.500	0.295	0.617
	薬学	0.523	0.393	**0.525**	0.400	0.475	0.435	0.286	0.310	0.348	0.415
	保健	0.495	0.468	**0.528**	0.490	0.536	0.510	0.209	0.289	0.282	0.440
	未来	0.553	0.524	**0.588**	0.545	0.566	0.545	0.341	0.333	0.235	0.276
	全学	**0.570**	0.536	0.523	0.528	0.529	0.528	0.172	0.394	0.064	0.441
下位 1/4 予測モデル	工学	0.455	0.508	0.522	**0.569**	0.539	0.474	0.099	0.439	0.338	0.436
	薬学	0.511	**0.531**	0.5	0.448	0.39	0.449	0.194	0.316	0.319	0.245
	保健	**0.482**	0.357	0.447	0.312	0.463	0.427	0.338	0.214	0.263	0.370
	未来	0.368	0.370	0.353	0.421	0.182	**0.444**	0.278	0.364	0.323	0.364
	全学	0.445	0.483	0.423	**0.510**	0.425	0.493	0.067	0.368	0.282	0.400

表 4 複数の機械学習手法における再現率の比較（ボールドはその行の最高値）

		勾配ブースティング決定木						決定木ベースではない手法			
		XGBoost		LightGBM		CatBoost		SVM		NN	
		②	②	①	②	②	②	③	③	③	②
下位 1/2 予測モデル	工学	0.636	0.766	0.628	0.774	0.612	0.815	0.372	0.750	0.364	**0.911**
	薬学	0.529	0.500	0.627	0.591	0.510	0.545	0.549	0.515	0.529	**1.000**
	保健	0.663	**1.000**	0.652	0.663	0.640	0.699	0.528	0.602	0.090	0.000
	未来	0.851	0.706	0.638	0.706	0.553	0.647	0.447	**0.882**	0.426	0.412
	全学	0.652	**0.827**	0.623	0.779	0.636	0.772	0.538	0.756	0.636	0.772
下位 1/3 予測モデル	工学	0.500	0.619	0.554	0.607	0.522	0.607	0.413	0.571	0.196	**0.690**
	薬学	0.531	0.324	0.656	0.351	0.438	0.405	0.250	0.243	0.375	**0.757**
	保健	0.473	0.393	0.509	0.446	0.473	0.446	0.127	0.196	0.182	**0.661**
	未来	0.448	0.524	0.517	**0.571**	0.517	**0.571**	0.241	0.238	0.138	0.190
	全学	0.490	0.525	0.438	0.520	0.433	**0.545**	0.096	0.318	0.034	0.384
下位 1/4 予測モデル	工学	0.352	0.533	0.423	0.617	0.437	0.533	0.056	0.483	0.310	**0.650**
	薬学	0.480	0.433	0.520	0.433	0.320	0.367	0.120	0.300	**0.720**	0.200
	保健	0.435	0.278	0.413	0.333	0.413	0.444	0.239	0.167	0.217	**0.611**
	未来	0.292	0.357	0.250	**0.571**	0.125	0.429	0.208	0.286	0.208	0.286
	全学	0.367	0.457	0.331	**0.479**	0.331	0.471	0.036	0.293	0.265	0.407

以上より，適合率を重視させた予測モデルによって，入学前の段階で，成績不良となるであろう学生を精確に絞り込む事は，特定の学生に集中したケアを早期に行う事ができる仕組みづくりのための重要な技術であり，現時点では，再現率こそ低いが，教育現場で試運用できる可能性も十分あると考察している．

また表 5 より，決定木ベースではない手法，特に，ニューラルネットワークは，再現率の最高値を多く取得している（ボールドのセルが多い）事が分かる．ニューラルネットワークでは，正則化技術（ドロップアウト、バッチ正規化など）や早期停止を用いる事で，過学習を防ぎつつ，汎化性能を高めることが可能であり，また，非線形活性化関

表 5 複数の機械学習手法における適合率の比較（ボールドはその行の最高値）

		勾配ブースティング決定木						決定木ベースではない手法			
		XGBoost		LightGBM		CatBoost		SVM		NN	
		①	②	②	②	②	②	④	②	④	②
下位 1/2 予測モデル	工学	0.796	0.617	**0.818**	0.644	0.840	0.620	0.738	0.633	0.770	0.661
	薬学	0.614	0.611	0.552	0.639	**0.667**	0.632	0.596	0.540	0.482	0.559
	保健	**0.663**	0.529	0.637	0.632	0.626	0.542	0.588	0.568	0.364	0.000
	未来	0.714	0.462	0.732	0.49	0.788	0.489	**0.778**	0.588	0.690	0.538
	全学	0.760	0.579	**0.767**	0.605	0.747	0.602	0.717	0.601	0.561	0.600
下位 1/3 予測モデル	工学	0.767	0.559	**0.823**	0.526	0.750	0.490	0.745	0.444	0.600	0.558
	薬学	0.515	0.500	0.438	0.464	**0.519**	0.469	0.333	0.429	0.324	0.286
	保健	0.520	0.579	0.549	0.543	**0.619**	0.595	0.583	0.550	0.625	0.330
	未来	0.722	0.524	0.682	0.522	0.625	0.522	0.583	0.556	**0.800**	0.500
	全学	0.680	0.547	0.650	0.536	0.682	0.512	**0.833**	0.516	0.700	0.517
下位 1/4 予測モデル	工学	0.641	0.485	0.682	0.529	**0.705**	0.427	0.400	0.403	0.373	0.328
	薬学	0.545	**0.684**	0.481	0.464	0.500	0.579	0.500	0.333	0.205	0.316
	保健	0.541	0.500	0.487	0.293	0.528	0.410	**0.579**	0.300	0.333	0.265
	未来	0.500	0.385	0.600	0.333	0.333	0.462	0.417	0.500	**0.714**	0.500
	全学	0.565	0.512	0.585	0.545	**0.591**	0.516	0.500	0.494	0.301	0.393

数を用いる事で，データの非線形性を効果的に捉えることができる．これらは，本稿のような複雑なデータセットに強く，モデルの再現率が担保されたとのだと考えられる．

さらに，表 4 より，CatBoost の手法が再現率の最高値を多く取得している（ボールドのセルが多い）事が分かる．CatBoost はカテゴリカルデータの処理に優れており，モデルの予測精度を向上させ，特に誤った正例の予測（偽陽性）を減少させる事で適合率を改善させる可能性を持つ．また，CatBoost はオーダーブースティングという技術を採用しており，データの順序に依存しない方法でモデルを学習するため，過学習を防ぎ，汎化性能を高める事ができる．本稿でのデータセットでも表 1 の通り，カテゴリカルデータが含まれており，それらが適切に処理された事で，結果的に，モデルが正例を誤って予測する事が減少し，適合率が改善したと考えられる．

5. まとめ

日本の大学においては退学者問題が深刻度を増してきている．学生が退学する理由の 1 つでもある「学業不振」を無くすためには，早期のうちから大学の新入生の成績を予測し，必要な学生に素早いフォローアップする体制が必要と言える．我々はこれまで，過去の学生に対する入学前のデータ（入試データ，入学前教育データなど）と入学後の成績の関係を機械学習し，新規で入学してくる大学生の 1 年後の成績を予測するシステムについて研究を行ってきた．

本稿では，複数の機械学習モデルを構築し，その予測性能の比較・議論してきた．結果として，予測モデルの F 値として 0.6～0.7 程度を取得する事ができた．今後は，モデルの解釈性についても議論していく．例えば，勾配ブースティング決定木では，各説明変数に対して，モデルの予測に与える影響の大きさを示す重要度が算出する事ができる．この重要度を用いて，成績に影響を与える要因を分析したり，成績不良者の早期発見を支援したりする手法についても検討していく．また，実際の教育現場での運用に向け，アプリデザインや機能等についても深く検討していく．

参考文献

[1] 紺田 広明，“退学に関わる支援策の現状と課題-自己点検・評価に記述される大学の実践から-，”福岡大学教育開発支援機構紀要, Vo.4, pp.40–49 (2022).
[2] T. Chen and C. Guestrin, “XGBoost: A Scalable Tree Boosting System,” Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp.785–794 (2016).
[3] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye and T. Y. Liu, “LightGBM: A Highly Efficient Gradient Boosting Decision Tree”, Proceedings of the 31st International Conference on Neural Information Processing Systems, pp.3149–3157 (2017).
[4] A. V. Dorogush, V. Ershov and A. Gulin, “CatBoost: Gradient Boosting with Categorical Features Support,” arXiv:1810.11363 (2018).
[5] Komei Arasawa, Shun Matsukawa, Nobuyuki Sugio, Hirofumi Sanada, Madoka Takahara, and Shun Hattori, “A Study on Forecasting Post-enrollment Grades of Admission Students to Universities Using Gradient Boosting Decision Tree”, Proceedings of the 26th International Conference on Human-Computer Interaction, LNCS (2024).