N-010

入学前教育の学修記録の解析に基づく大学入学後の GPA 予測
Forecast of University Student's Grade Point Average based on Analyzing Study Logs in Pre-admission Education

荒澤 孔明‡　松川 瞬‡　杉尾 信行‡　真田 博文‡
Komei Arasawa　Shun Matsukawa　Nobuyuki Sugio　Hirofumi Sanada

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]．本研究では，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特徴量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$1/N$になるか否かを予測するモデルの構築を行い，その予測性能や予測に重要な説明変数を明らかにする．

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 コースそ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 CatBoost[2] を用いて予測を行う．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．

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

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

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．

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

3. 評価実験

本稿では，大学入学者の 1 年後の GPA の順位区分が下位 1/2 になるか否かを予測するモデル，下位 1/3 になるか否かを予測するモデル，下位 1/4 になるか否かを予測するモデルの 3 つを構築し，それぞれの予測性能の評価を行う．なお，学習には，2021 年度の新入生データを，またテストには，2022 年度の新入生データを用いた．各年度の新入生の 1 年後の成績の分布を表 1 に示した．

表 3～表 6 に，3 つのモデルの予測性能を示した．再現率を見ると，下位 1/2 を成績不振の学生とした時，そのうち 6 割を，下位 1/3 を成績不振の学生とした時，そのうち 4 割を，下位 1/4 を成績不振の学生とした時，そのうち 3 割を事前に捉えるができるモデルであると分かった．予測性能の及第点は，教育現場に依っても異なるため，今後は，大学教員らへのヒアリングなどを通じ，これらの予測性能に関してより深い評価を行っていく．

表 1　各年度における新入生の 1 年後の成績区分

	目的変数	
2021 年度入学生（学習データ）	下位1/2	301（54.8%）
	それ以外	248（45.2%）
	下位1/3	200（36.4%）
	それ以外	349（63.5%）
	下位1/4	156（28.5%）
	それ以外	392（71.5%）
2022 年度入学生（テストデータ）	下位 1/2	316（54.9%）
	それ以外	260（45.1%）
	下位1/3	208（36.1%）
	それ以外	368（63.8%）
	下位1/4	166（28.8%）
	それ以外	410（71.1%）

表 2　大学合格者の入学後の成績予測を行うために利用した 16 種類の説明変数

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

表 3　下位1/2の予測モデルにおける混同行列

	Fcst. Low 1/2	**Fcst. Others**
Act. Low 1/2	200	116
Act. Others	68	192

表 4　下位1/3の予測モデルにおける混同行列

	Fcst. Low 1/3	**Fcst. Others**
Act. Low 1/3	90	118
Act. Others	42	326

表 5　下位1/4の予測モデルにおける混同行列

	Fcst. Low 1/4	**Fcst. Others**
Act. Low 1/4	55	111
Act. Others	37	373

表 6　各タスクにおける予測性能

	再現率	**適合率**	**F 値**
1 年後の成績が下位 1/2 学生の予測	0.633	0.746	0.685
1 年後の成績が下位 1/3 学生の予測	0.433	0.682	0.529
1 年後の成績が下位 1/4 学生の予測	0.331	0.598	0.426

また，表 7 には，予測モデルの中で算出された各説明変数の重要度を示した．この結果，入学後の成績に影響を与える変数は「高校在籍時の平均評定値」や「入学前教育ドリルの実施回数」さらに「基礎学力テストの合計得点」である事が分かり，この結果は，直感と一致するもの窺える．

4. 今後の課題

本稿では，CatBoost に基づき，大学合格者の入学後の成績を予測するモデルの構築を行ってきた．CatBoost は，勾配ブースティング決定木の 1 つであり，他にも，XBGoost や LightGBM といったプラットフォームがある．各々の予測性能はタスク依存とされているため，今後は，これらの手法を用いた予測モデルに関しても深く議論を行っていく．

表 7　CatBoost で算出された各説明変数の重要度

	下位 1/2 予測モデル	**下位 1/3 予測モデル**	**下位 1/4 予測モデル**
1	0.000	3.881	1.329
2	0.035	2.279	0.952
3	1.136	6.238	5.669
4	**32.774**	**11.496**	**17.128**
5	3.474	7.632	3.008
6	1.570	7.380	5.399
7	0.763	5.898	3.198
8	0.590	5.334	7.255
9	4.216	5.726	7.766
10	**12.350**	**5.403**	**8.610**
11	2.914	6.070	3.831
12	1.289	5.221	3.474
13	**11.888**	**6.440**	**8.156**
14	1.884	6.861	4.935
15	**17.850**	**9.302**	**11.205**
16	7.266	4.839	8.084

参考文献

[1] 白鳥 成彦, 大石 哲也, 田尻 慎太郎, 森 雅生, 室田 真男, “予測 GPA の推移を用いた 1 年次春学期学修状態の類型化,” 教育システム情報学会誌, Vol.39, No.4, pp.440-451 (2022).

[2] A. V. Dorogush, et.al., “CatBoost: Gradient Boosting with Categorical Features Support,”, arXiv:1810.11363 (2018).