P1-19

修学ビッグデータを用いた退学および留年リスクの予測と支援への応用

Prediction of Dropout and Grade Retention Risks Using Educational Big Data and Its Application for Student Support

吉本 悠真[*1], 鈴木 崇太朗[*1], 山本 知仁[*1],
Yuma YOSHIMOTO, Sotaro SUZUKI, Tomohito YAMAMOTO
[*1] 金沢工業大学大学院工学研究科情報工学専攻
[*1]Graduate School of Engineering, Kanazawa Institute of Technology
Email: c6500676@st.kanazawa-it.ac.jp

あらまし：文部科学省の調査によると，令和 5 年度における全国の大学・短期大学の退学率は 2.1%とされている．金沢工業大学では，年間退学率が 3%前後で推移しており，退学の抑制に向けた具体的な対策が必要とされている．そこで本研究では，大学のデータベースに蓄積された学生の修学状況に関するデータを用いて，退学および留年を予測するモデルを構築した．また，モデルの予測結果に SHAP を適用することにより，各学生のリスク要因に応じたフィードバックメッセージの提示を実現した．
キーワード：教学 IR，データサイエンス，AI，学習支援

1. はじめに

近年の日本の教育現場においては，大学進学率が増加傾向にある．その一方で，退学などの問題により，大学での学びを継続できなくなる学生も少なくはない．文部科学省の調査によると，令和 5 年度に大学および短期大学を退学した学生は 56,710 名にのぼり，退学率は 2.1%である(1)．金沢工業大学（以下，金沢工大）においては，近年の年間退学率が 3%前後で推移しており，退学の抑制に向けた具体的な対策が必要になっている．

このような背景のもと，教学 IR の一分野として，機械学習を活用した退学者予測に関する研究が行われている．たとえば，近藤らの研究では，入学前学習の提出状況や 1 年次前期第 5 週までの出席状況を用いることで，約 40%の精度で 3 年次開始時の在籍状況が予測可能であることが示されている(2)．また，石川らの研究では，1 年次の GPA や取得単位数，高等学校の偏差値などを用いることで，約 60%の精度で退学を予測できることが示されている(3)．金沢工大では，寺井らの研究で退学者予測モデルの構築が行われてきた(4)．寺井らの研究では，成績が確定した学期の 2 学期先までの退学を予測するモデルが提案され，モデルの予測結果に基づいたフィードバックが実施されている．しかし，この研究で提案されたモデルには，1 年次前期における退学者を見逃してしまうという問題点があった．また，退学者のみをターゲットとしており，留年者については支援の対象としていなかった．

本研究ではこれらの問題に対応するために，LMS のアクセスログや出席率などの随時更新されるデータを用いて，1 年次前期中の予測を可能とするモデル（以下，早期予測モデル）を構築する．さらに，従来の成績および出席情報に基づくモデルを拡張し，LMS のアクセスログや課題提出状況などの多様なデータを取り入れた新たなモデル（以下，学期ごと予測モデル）を構築する．いずれのモデルについても，留年者も考慮したモデル設計にすることで，大学生活への適応に課題を抱える学生をより包括的に特定することを目指す．

2. 退学者と留年者に関する予備的な解析

金沢工大では，2020 年より学内の教育 DX を推進する取り組みの一環として大学のデータベースを統合し，解析の基盤を構築している．本研究では，このデータベースにある修学状況に関するデータを解析の対象とする．

予測モデルを構築するにあたって，退学者および留年者に関する解析を行った．結果として，1 年次の退学者が最も多く，学年が進むにつれて退学者数が減少する傾向がみられた．また，学力不足や修学意欲の低下によって退学に至る学生は高年次生に多いことや，他校への進路変更によって退学に至る学生は低年次生に多いこと，経済的理由によって退学に至る学生は学年に関係なく一定の割合で存在することなどが明らかとなった．

2016 年度から 2021 年度の入学生を対象に，留年経験と退学率の関係についても解析を行った．結果として，留年を経験した学年が低いほど退学に至る割合が高い傾向が見られた．特に，1 年次または 2 年次に留年を経験した学生では，卒業に至った割合は全体の約 2 割にとどまり，退学に至る割合が顕著に高いことが明らかとなった．

3. 退学および留年予測モデル

3.1 モデルの概要

金沢工大における 2018 年度から 2022 年度入学生のデータを用いて学習を行い，2023 年度入学生のデータを用いて評価を行った．また，クラスタリングとアンダーサンプリングを組み合わせることにより，クラス不均衡問題に対処しつつ，学生の多様な特性

表 1 早期予測モデルの予測精度

Rank	Model	Accuracy	Precision	Recall	F1 Score
1	Multilayer Perceptron	0.91	0.42	0.61	0.50
2	Random Forest Classifier	0.90	0.38	0.66	0.48
3	Gradient Boosting Classifier	0.90	0.38	0.62	0.47
4	Extra Trees Classifier	0.90	0.39	0.60	0.47
5	Decision Tree Classifier	0.88	0.37	0.58	0.45

表 2 学期ごと予測モデルの予測精度

Rank	Model	Accuracy	Precision	Recall	F1 Score
1	Random Forest Classifier	0.89	0.32	0.75	0.45
2	Extra Trees Classifier	0.88	0.32	0.71	0.44
3	CatBoost Classifier	0.88	0.31	0.72	0.43
4	Blend model（rf + et）	0.88	0.32	0.66	0.43
5	TabPFN	0.87	0.29	0.69	0.41

を反映した代表性の高いデータセットを構築し，モデルの学習を行った．また，Grinsztajn らによると，数千行規模の表形式データの予測に関しては，ツリーベースモデルが最も優れた精度を示す傾向にあると報告されている(5)．そのため，複数のツリーベースモデルを効率的に比較することができる PyCaret を用いてモデルの学習および評価を実施した．

3.2 早期予測モデルの結果

評価用データに対する予測精度が良好であったモデルの結果を表 1 に示す．表 1 より，多層パーセプトロンやツリーベースモデルが比較的良好な性能であることがわかる．そして，1 年次前期の第 8 週までのデータを用いて，約 50%の精度で 1 年次終了までに発生する退学および留年を予測できることが明らかとなった．

3.3 学期ごと予測モデルの結果

評価用データに対する予測精度が良好であったモデルの結果を表 2 に示す．表 2 から，ツリーベースモデルおよび複数のモデルを組み合わせたブレンドモデルが，比較的高い性能を示していることが確認できる．予測を実施した学期により精度には一定のばらつきが見られたものの，約 45%の精度で 2 学期先までの退学および留年を予測できることが明らかとなった．寺井らが提案したモデルと比較して約 5%の精度向上が見られたため，学習成果に加えて行動データや属性データも考慮することが，退学および留年の予測において有効である可能性が示唆された．

4. フィードバックメッセージの提示

金沢工大では，寺井らの研究により 2022 年度から，退学リスクがあると予測された学生に対し，フィードバックを提示する取り組みが実施されている．具体的には，退学リスクがあると予測された学生に対して，SHAP を適用し，リスクを高めている要因を特定した上で，その結果に基づくフィードバックを提示している．SHAP とは，ゲーム理論におけるシャープレイ値の概念を応用し，各特徴量が予測結果に与える影響度を定量的に算出する手法である．

従来は，数理科目・専門科目・出席といった観点から学生を 4 つのカテゴリに分類し，それぞれに応じたフィードバックを提供していた．本研究では，成績や出席に加えて，LMS のアクセスログや課題提出状況といった行動データも活用することで，成績・出席・アクセスログ・課題提出の 4 項目を主な分類軸とし，さらに成績と出席については科目群ごとに細分化した．これにより，各学生の修学状況に即した，より的確なフィードバックの提供を実現した．本研究で実施したフィードバックは，2025 年 5 月現在，学内のポータルサイトを通じて学生に提示されている．

5. まとめと今後の展望

本研究では，退学および留年予測モデルを構築した．いずれのモデルも 4-5 割の精度で退学および留年を予測することできた．また，SHAP を活用することで各学生のリスク要因を特定し，その結果に基づいたフィードバックを提示することができた．

今後は，本研究にて未活用の自然言語データなどを取り入れた場合の予測精度を検証するとともに，フィードバックを受け取った学生の修学状況の変化を解析し，長期的な効果を検証する予定である．

参考文献

(1) 文部科学省:”令和 5 年度 学生の中途退学者・休学者の調査結果について”, https://www.mext.go.jp/content/20240627-mxt_gakushi01-000013028_1.pdf （参照 2025.4.1）
(2) 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”, 教育システム情報学会誌, Vol.33, No.2, pp.94-103 （2016）
(3) 石川千温, 石本翔真:“機械学習を用いた退学予測に基づくエンロールメントマネジメントシステムの構築”, 情報処理学会論文誌デジタルプラクティス, Vol.4, No.2, pp.1-8 （2023）
(4) 寺井朝人, 鈴木崇太朗, 山本知仁: “修学ビッグデータを用いた退学者予測と学習支援”, 2023 年電気・情報関係学会北陸支部連合大会予稿集, p.177 （2023）
(5) Grinsztajn, L., Oyallon, E. and Varoquaux, G.:“Why do tree-based models still outperform deep learning on typical tabular data?”, Advances in Neural Information Processing Systems, Vol.35, pp.507-520 （2022）