

[原著論文：査読付]

機械学習を用いた出席率からの中途退学者予測モデルの構築

島尻 芳人*, 堂野崎 融*

Dropout Prediction Modeling from attendance rate Using Machine Learning

Yoshito SHIMAJIRI*, Tooru DOUNOSAKI*

要 旨

18歳人口の減少が影響を及ぼし、大学への進学者数が減少すると予測されている。そのため、私立大学間の学生確保競争が激化し、大学法人の経営は今後一層の困難が予想される。さらに、中途退学の低下を通じて、中途退学の予防に取り組むことは、授業料収入の減少を防ぐだけでなく、受験生に対しても魅力的な大学としてのイメージを向上させる助けとなることが期待される。そこで、本研究では、機械学習の手法の1つであるランダムフォレスト(Extra Trees Classifier)を用いて、15週間分の出席率情報から中途退学者の予測モデルの構築を試みた。結果、約90%という高い精度で予測をすることに成功した。さらに、5週間分の出席率情報のみを用いた場合は、予測精度が約80%であることが明らかになった。これらの成果は、今後の大学運営方針の策定や学生支援体制の強化に向けて、有益な示唆を提供するものと期待される。

キーワード：機械学習, Institutional Research, 中途退学予防

Abstract

The number of students entering universities is expected to decrease due to the decrease in the number of 18-year-olds. As a result, competition among private universities to secure students is expected to intensify, making the management of university corporations among difficult in the future. In addition, it is expected that the prevention of dropout through the reduction of dropout will not only prevent the decrease of tuition revenue but also help to improve the image of the university as an attractive university for prospective students. Therefore, in this study, we attempted to construct a prediction model of dropouts from 15 weeks of attendance rate using a random forest (Extra Trees Classifier), which is one of the machine learning methods. As a result, we succeeded in predicting the dropout students with a high accuracy of ~90%. Furthermore, when only 5 weeks of attendance rate were used, the prediction accuracy was found to be ~80%. These results are expected to provide useful suggestions for the formulation of future university management policies and the enhancement of student support systems.

KEY WORDS : Machine Learning, Institutional Research, Dropout Prevention

*九州共立大学経済学部

*Faculty of Economics, Kyushu Kyoritsu University

1. はじめに

日本の18歳人口は、2018年以降減少傾向にある。一方で、大学・短期大学への進学率(過年度高卒者などを含む)は、2008年から2021年にかけて55%から59%に増加した¹⁾。そのため、大学進学者数は横ばいであったが、今後は減少に転じ、2040年までには、大学進学者数が現在より20%減少することが見込まれる²⁾。その結果、多くの大学で定員割れによる授業料収入の減少が生じ、大学法人の経営が一層困難になると予測される³⁾。このような状況に対処するため、大学は内部の様々なデータを分析する取り組みを行っている。教育、経営、財務情報を含む大学内データの分析を通じて、大学の戦略計画(経営 IR, Institutional Research)や教育の改善(教学IR)を行う活動が、多くの大学で活発に展開されている⁴⁾。

中途退学数の増加は、大学経営において主に2つのデメリットが考えられる。まず1つ目は、授業料収入の減少によって大学経営が圧迫されることである。2つ目は、教育機関としての信頼を損ない、入学希望者の減少につながる可能性があることである。中途退学を予防するためには、中途退学の可能性がある学生を早期に発見し、適切なケアを行うことが重要である。そのため、島尻・堂野崎(2023)⁵⁾は、2016年度から2019年度に入学した学生に対して行った新入生調査(JSAAP, Joint Student Achievement Assessing Project⁶⁾)と2016年度から2019年度の入学生の中途退学者情報をもとに、機械学習のテクニックの1つであるランダムフォレスト(Extra Trees Classifier⁷⁾)を用いて、中途退学者予測モデルを構築した。このモデルの予測精度は約6割であった。新入生調査は、入学前の生活習慣などに関する約100問の設問からなる。つまり、大学入学後の学習意欲や成績が考慮されていない。そのため、予測精度が高くなかったと考えられる。

そこで、本研究では、学習意欲が大学の授業出席率に反映されていると仮定し、各学期の各週の出席率情報を活用して中途退学者を高い精度で予測できるか検証することを目的とする。2章では、使用したデータおよびデータクリーニングについてまとめる。3章では、出席率と中途退学の関係について述べる。4章では、機械学習モデルの構築および評価について述べる。5章では、構築した機械学習モデルを実際の大学運営で活用するために必要な改善点を考察する。最後に6章にて、本研究をまとめる。

2. データ及びデータクリーニング

本研究では、2016年度から2019年度に入学した学生の各学期15週間分の出席率情報と中途退学の有無情報を用いた。どちらのリストも個人情報保護の観点から、データを取得後、氏名・ふりがな・学籍番号を入学年の西暦下2桁とランダムな5桁の数字を繋げた7桁の数字に変換した。この変換を行うことで、個人情報を特定できない状態で分析を行った。

出席率情報は、各学生の週ごとの平均出席率が集計されている。2020年度以降の出席率情報は、その週までの累計の出席率がまとめられており、集計方法が異なるため、本研究では使用しなかった。中途退学者情報には、退学日とともに進路変更、経済的理由、就学意欲の低下、病気などの退学理由が簡潔にまとめられている。ランダムに変換した7桁の数字を使い、出席率情報と中途退学者を照合し、全学生の出席率に中途退学の有無情報を付与した。

3. 出席率と中途退学の関係

中途退学には、進路変更、金銭的理由、学習意欲の低下など、さまざまな理由がある。他大学受験や海外留学などの進路変更は積極的な動機であり、金銭的理由による中途退学や結婚・災害・家族の介護などによる中途退学は、やむを得ない事情である。一方で、学習意欲の低下は、学生と大学の担当教員との面談を通して原因を探り、解決策を見出すことができる可能性がある。図1は、2016年度から2019年度に入学した学生の中で中途退学した学生の中途退学理由の内訳を示している。この学内データにおいて、就職への進路変更(26.9%)、就学意欲低下(23.1%)、その他(17.8%)、経済的理由(14.0%)、他大学への進路変更(9.6%)、専門学校への進路変更(5.3%)、病気(2.4%)、学力不足(0.4%)、懲戒処分(0.4%)が中途退学の理由として挙げられている。

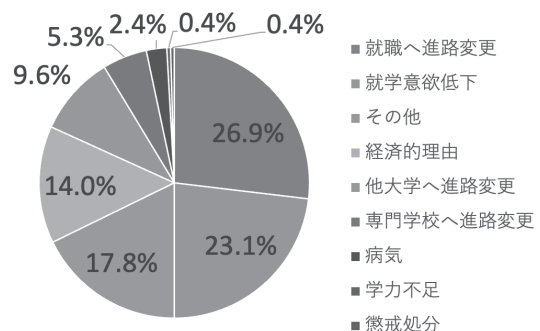


図1: 2016年度から2019年度に入学した学生のうち中途退学をした学生の中途退学理由の内訳

図2は、退学理由を3つのカテゴリー ((a)進路変更による中途退学, (b)学習意欲の低下による中途退学, (c)その他の理由による中途退学)に分類し、各カテゴリーに属する学生の1学期分(15週間分)の出席率の推移を示した図である。赤線は中途退学しなかった学生の出席率の推移を示しており、6週目、7週目、14週目で出席率が約5-10%低下するが、75%以上の学生が出席率80%以上を維持していることが確認できる。一方、①進路変更による中途退学のカテゴリーに属する学生の場合、1週目から3週目までは80%以上の出席率を維持しており、中途退学しなかった学生と大きな差は見られない。しかし、その後、徐々に出席率が低下し、最終週には、10%程度の出席率にまで減少する。②学習意欲の低下による中途退学のカテゴリーに属する学生の場合、1週目から出席率が50%を下回っており、9週目には、ほぼ0%となっている。③その他の理由による中途退学のカテゴリー、すなわち、経済的理由、就職、病気等を理由に中途退学した学生の場合、①と②のカテゴリーの中間的な出席率の推移となっている。

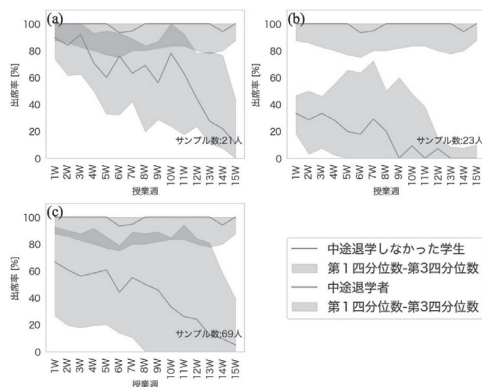


図2: (a)進路変更を理由に中途退学した学生の15週間の出席率の推移, (b)学習意欲の低下等を理由に中途退学した学生の15週間の出席率の推移, (c)その他の理由(就職、病気等)を理由に中途退学した学生の15週間の出席率の推移。中途退学しなかった学生の15週間の出席率の中央値が実線、第1四分位数から第3四分位数の範囲が赤で示されている。パネル(a), (b), (c)において、各理由で中途退学した学生の15週間の出席率の中央値が実線、第1四分位数から第3四分位数の範囲が灰色で示されている。

九州共立大学キャリア支援課では、3週目の出席率が50%未満の学生に対して、中途退学を予防するために個人面談をゼミ担当員に依頼している。この基準による面談対象の選定は、2つ目のカテゴリーの学習意欲の低下により中途退学した学生の大部分を選定できていることになる。ただし、この基準による選定では、カテゴリー①の進路変更で中途退学した学生の大部分と、カテゴリー③のその他の理由で中途退学した学生の半数は、この基準では選定できていないことを示している。学習意欲の低下による中途退学の予防は、進

路変更や経済的な理由による中途退学を防ぐことより、大学としての役割が大きいと考えられる。そのため、現行の3週目の出席率が50%未満という基準は適切な基準であったと考えられる。

4. 機械学習モデルの構築および評価

九州共立大学では、3週目の出席率が50%未満を基準として、中途退学を予防するための個人面談の対象者を選定している。しかし、前章で述べた通り、この基準では、半数以上の中途退学者を抽出できていない。そこで、中途退学の可能性が高い学生をより高い精度で選定するため、本章では、機械学習を活用して、中途退学者の予測モデルを構築する。

4-1. 使用データ

3章で示した通り、中途退学のケースにおいても、中途退学理由により出席の推移が異なることが明らかになった。しかし、機械学習を用いて予測モデルを構築する際に、中途退学理由ごとに教師データとなる中途退学者の出席率情報を分けてしまうと、学習に必要な十分なサンプル数(中途退学者数)を得ることができない。そのため、本研究では、中途退学理由ごとの分類は行わず、中途退学者全般を対象として予測モデルを構築することとする。

4-2. 15週間分の出席率情報からの予測のモデル構築とその精度

本研究では、Python モジュールの`pycaret`⁸⁾を用いた。具体的には、以下の手順に従ってモデルを構築した。特徴量として15週間の各週の出席率を用いて、教師データとしては中途退学の有無情報を用いた。

1. `compare_models` 関数を用いて、本研究において最もよいモデルの選択

`pycaret` では、Extra Trees Classifier (ET), Random Forest Classifier (RF), Gradient Boosting Classifier (GBC), Light Gradient Boosting Machine (Light GMB)などの18のモデルが利用可能である。 k -hold cross-validation法で各モデルのパフォーマンスを評価する。その中から、決定係数が最も高いモデルを最も良い結果のモデルとして選択する。4-2章では、Extra Trees Classifier (ET)を選択した。

2. 選択したモデルのハイパーパラメータの最適化

`create_model` 関数を用いて、選択したモデルでモデルを作成し、`tune_model` 関数を用いて、選択したモデルのハイパーパラメータの最適化する。

3. モデル精度の評価

`plot_model` 関数を用いて、ハイパーパラメータが最適化されたモデルのパフォーマンスを評価する。 k -hold cross-validation法で、 k 分割されたサンプルセットに対して、モデル構築に使われなかった1サンプルセットを使って、モデル精度の評価を行った。本研究では、1サンプルセットは67名に対応する。

図3(a)は、実際に中途退学しなかった学生と中途退学した学生の中途退学の有無の予測結果を示した図である。実際に中途退学しなかった学生31名のうち、26名(84%)が中途退学しないと予測され、5名(16%)が中途退学すると予測された。一方で、実際に中途退学した36名のうち、3名(8%)が中途退学しないと予測され、33名(92%)が中途退学すると予測された。

図4(a)はモデルのPrecision(精度), Recall(再現率), f1(f1 score), Supportを示している。Precision(精度)は、以下のように定義される。

$$Precision = \frac{True\ Positives\ (TP)}{True\ Positives\ (TP) + False\ Positives\ (FP)} \cdots (1),$$

True Positivesは、実際に中途退学(非中途退学)であり、中途退学(非中途退学)と予測されたサンプルの数を示す。また、False Positivesは、実際に非中途退学(中途退学)であり、中途退学(非中途退学)と予測されたサンプルの数を示す。したがって、Precision(精度)は、モデルが中途退学と予測したケースのうち、実際に中途退学であったサンプルの割合を測る指標であり、

モデルの予測精度を示す。中途退学者のPrecisionは0.917に対して、中途退学しなかった学生のPrecisionは0.839であった。Recall(再現率)は、以下のように定義される。

$$Recall = \frac{True\ Positives\ (TP)}{True\ Positives\ (TP) + False\ Negative\ (FN)} \cdots (2),$$

False Negativeは、実際に中途退学したにもかかわらず、中途退学と予測されたサンプルの数を示す。つまり、Recallの値が高いと正しく中途退学者を予測できており、見逃すリスクが低いことを示す。中途退学者のRecallは0.868に対して、中途退学しなかった学生のRecallは0.897であった。f1 (f1 score)は、PrecisionとRecallの調和平均として計算される。これは、PrecisionとRecallのバランスを示す指標として使われる。0～1の間の値をとり、1に近いほど、モデルの性能が良いことを示す。このf1は以下のように定義される。

$$f1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \cdots (3).$$

中途退学者のf1は0.892に対して、中途退学しなかった学生のf1は0.867であった。Supportは、各クラスの実際のサンプル数を示す。

モデルは、中途退学者および中途退学しなかった学生の両方をよく予測している結果を示している。しかし、中途退学者を予測する際に、確信度が高いものの、実際の中途退学者の中で一部を見逃している可能性があるため、さらなる精度の向上が求められる。以上のように、Extra Trees Classifier を用いて、15週間分の出席率をもとに中途退学者の予測モデルを構築することに成功した。予測精度は、8-9割と高い精度を達成している。

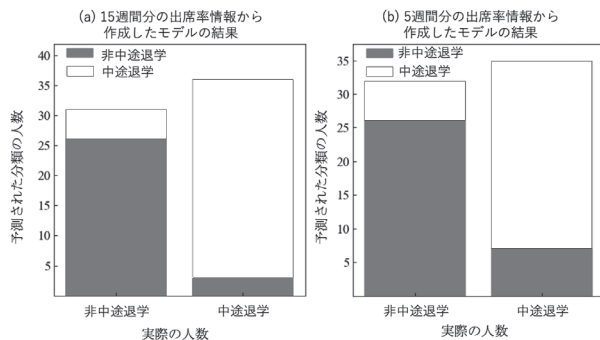


図3 : (a) 15週間分と(b)5週間分の出席情報から作成したモデルの実際に中途退学しなかった学生と中途退学した学生の中途退学の有無の予測結果。濃い灰色は、中途退学しないと予測された学生の数。薄い灰色は、中途退学すると予測された学生の数。

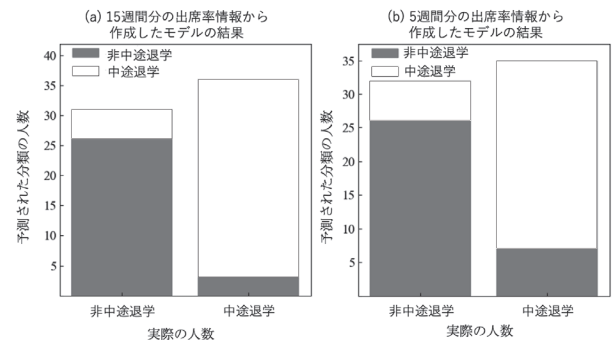


図4 : (a) 15週間分と(b)5週間分の出席情報から作成したモデルの Precision(精度), Recall(再現率), f1(f1 score), Support.

4-3. 5週間分の出席率情報からの予測のモデル構築とその精度

本研究では、15週間分の出席率の推移を学習することで、高い精度(約90%)で中途退学の可能性を予測できることを示した。つまり、学期終了後に中途退学者を高い確率で予測することが可能である。しかし、中途退学者予測モデルを中途退学の予防に実際に活用するためには、学期の前半に中途退学の兆候がある学生を特定し、予防策を検討する必要がある。そこで、ここでは、学期の最初の5週間分の出席率の情報のみから、4-2章と同様に中途退学者予測モデルを構築し、その精度を評価する。

4-2章と同様の手順でモデルを作成し、その結果を図3(b)および図4(b)に示した。図3(b)で示したように、実際に中途退学しなかった学生32名のうち、26名(81%)が中途退学しないと予測され、6名(18%)が中途退学すると予測された。これは、15週間分の出席率情報を用いて作成したモデルと比べ2-3%精度が悪化している。一方で、実際に中途退学した35名のうち、7名(20%)が中途退学しないと予測され、28名(80%)が退学すると予測された。これは、15週間分の出席率情報を用いて作成したモデルと比べ10%精度が悪化している。図4(b)で示したように、Precision(精度), Recall(再現率), f1(f1 score)も15週間分の出席率情報を用いて作成したモデルと比べ数%から10数%悪化していることが明らかになった。

5週間分の出席率情報から中途退学者予測モデルを構築すると、予測精度は80%程度であり、15週間分の出席率情報から中途退学者予測モデルと比べ、約10%悪化することが明らかになった。

5. 大学運営に活用するために必要な改善点

本研究では、15週間分の出席率の推移を学習することで、高い精度(約90%)で中途退学の可能性を予測できることを示した。つまり、学期終了後に中途退学者を高い確率で予測することが可能である。さらに、5週間分のみ出席率情報からでも約80%の精度で中途退学者を予測できることを示した。しかし、中途退学者予測モデルを中途退学の予防に実際に活用するためには、学期の前半に中途退学の兆候がある学生をより高い精度で特定し、予防策を検討する必要がある。先行研究(島尻・堂野崎 2023)で行った大学入学前の情報が詰まった新入生調査(JSAAP)の情報と本研究で用

いた出席率の情報の両方の情報を使いモデルを構築することで、早期に高い予測できるようになることが期待される。

また、本研究では、機械学習に必要な十分なサンプル数(中途退学者数)を確保するため、異なる出席率の推移を見せる異なる中途退学理由を1つにまとめ予測モデルを構築した。しかし、中途退学理由に応じて予測できるようになれば、中途退学の予防策の実施が容易になることが期待される。今後、十分なサンプル数が得られた場合、中途退学の有無だけでなく、中途退学の理由も予測できるように改善する予定である。

6. まとめ

本研究では、2016年度から2019年度に入学した学生の各学期15週間分の出席率と中途退学の有無情報を用いて、中途退学者の予測に関する機械学習モデルの構築を行った。本研究の結果を以下にまとめる。

1. 15週間の出席率の推移は、中途退学しなかった学生、進路変更のため中途退学した学生、学習意欲の低下のため中途退学した学生、その他の理由で中途退学した学生の4つのグループで異なることが明らかになった。進路変更のため中途退学した学生は、初めは80%の高い出席率であるが、徐々に出席率が低下していく。学習意欲の低下が原因で中途退学した学生は初めから出席率が50%未満であり、9週目以降にはほぼ0%の出席率にまで低下する。その他の理由で中途退学した学生は、進路変更と学習意欲の低下を理由に中途退学した学生の中間の出席率の推移であった。

2. 九州共立大学が中途退学予防のために行っている個別面談の対象学生の選定基準である3週目に出席率50%未満という基準は、学習意欲の低下のため中途退学した学生の大部分を抑えている。しかし、進路変更や経済的理由による中途退学の場合、この基準では十分に対象となる学生を選び出すことができないことが明らかになった。

3. 15週間分の出席率の推移をもとに、機械学習のテクニックの1つであるExtra Trees Classifierを使って、中途退学者の予測モデルを構築した。結果、中途退学者を高い精度(約9割)で予測可能であることが明らかになった。

4. 中途退学者予測モデルを中途退学の予防に実際に活用するためには、大学入学前の情報が詰まった新入生調査(JSAAP)の情報と出席率の情報の両方の情報を

使いモデルを改善することで、早期に高い予測できるようにする必要がある。

謝辞

本研究のデータ収集にあたり、九州共立大学キャリア支援課 池本 友洋氏および古川 裕貴氏にご協力をいただいた。この場で感謝を申し上げる。

参考文献

- 1) 「学校基本調査」(文部科学省)(2022) : [Online]. Available: [https://www.mext.go.jp/b_menu/toukei/chousa01/kihon/1267995.htm]
- 2) 中央教育審議会大学分科会将来構想部会 (2022) : [Online]. Available: [https://www.mext.go.jp/b_menu/shingi/chukyo/chukyo0/toushin/1411360.htm]
- 3) 日本私立大学教会アルカディア学法 (n. d.) : [Online]. Available: [<https://shidaikyo.or.jp/riihe/research/arcadia/0585.html>]
- 4) 小林雅之・山田礼子 (2016) : 大学の IR 意思決定支援のための情報収集と分析. 慶應義塾大学出版.
- 5) 島尻芳人・堂野崎融 (2023) : 機械学習による公務員合格者および中途退学者の予測モデルの構築. 九州共立大学学術情報センター, 6(15-24).
- 6) ジェイ・サーブ (n. d.) : [Online]. Available: [<https://jsaap.jp/>]
- 7) Geurts P., Ernst D., Wehenkel L. (2006) : Extremely randomized trees. Machine Learning, 63(3).
- 8) Moez Ali (2020) : PyCaret: An open source, low-code machine learning library in Python. [Online]. Available: [<https://www.pycaret.org>]

Received date 2023年10月17日

Accepted date 2024年1月10日