様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　６ 年　 ５ 月 ２９ 日現在

機関番号：３０１０３
研究種目：基盤研究(C)（一般）
研究期間：2019～2023
課題番号：１９Ｋ０２８６８

研究課題名（和文）クラウドによる機械学習を利用したエンロールメントマネジメントシステムの構築

研究課題名（英文）Building an Enrollment Management System Using Machine Learning in the Cloud

研究代表者
石川　千温（Ishikawa, Chiharu）

札幌学院大学・経済経営学部・教授

研究者番号：９０２８５４９５

交付決定額（研究期間全体）：（直接経費）　　1,900,000円

研究成果の概要（和文）：大学におけるIR分析を発展させ，エンロールメントマネジメントに不可欠な学生の学修状況の把握と分析，とりわけ，退学へと至る問題状況を早期に察知し，予測するためのシステムを機械学習の技術を用いて構築した．
機械学習に親和性の高いPython言語とExcelを用いて開発したこの退学予測システムでは，卒業年が2022年である学生の退学予測において，実践上有効と思われる精度で退学を予測することができ，本システムの可能性が実証された．また，これら分析システムをクラウドサービスを用いて所属機関内で共有することを可能にし，実用上の成果を得た．

研究成果の学術的意義や社会的意義
大学全入化に伴い大学（特に私立大学）の中途退学者の増加は社会的課題になっており，それを防ぐ取り組みが各大学に求められている．一方で，学生の様々な学修データや行動履歴などを一元化して，その状況を可視化するIR（Institute Research)は，まだ，分析結果の可視化のレベルに留まっており，退学者防止など実用上の対策に結びついていない．そこで，これらIRデータを単なる可視化に留めず，機械学習（AI)による退学予測システムに用いることで，大学の退学者と未然に防ぐ取り組みに用いることができるようになる．

研究成果の概要（英文）：We have developed a system for understanding and analyzing students' academic progress, which is indispensable for enrollment management, by developing IR analysis in universities. In particular, we used machine learning technology to construct a system for early detection and prediction of problematic situations that lead to withdrawal from the university. The system was developed using the Python language, which has a high affinity for machine learning, and Excel. The system was able to predict the withdrawal of students whose graduation year is 2022 with an accuracy considered to be effective in practice, demonstrating the potential of the system. The system can be shared among institutions using cloud services, and practical results were obtained.

研究分野：教育工学

キーワード：IR　エンロールメントマネジメント　退学予測　機械学習　クラウドサービス

１．研究開始当初の背景

研究開始当初，申請者は所属大学の教学 IR 委員会のメンバー（委員長）として，大学の様々なデータを収集，整理し分析を行っていた．それらの分析結果は学外へはほとんど公開されることなく，大学執行部の経営判断や各部局の事業計画に活かされることのみ使用されてきた．分析結果の大半は，学生の入学前から入学時，在籍時，卒業時の情報を串刺しにして，いくつかの要素間の関連性（相関）を見出し，統計的見地から報告するものであり，基本的には結果論であって，そこから将来に向けて予測を行うものではない．しかしながら，本学において，教学 IR の導入時には EM（Enrollment Management）もその一つの目的として掲げられており，これまで分析した結果から何らかの予測モデルを導入して将来の学生の修学指導，支援に用いることが求められていた．残念ながらその実現は，現状では非常に困難な状況であった．

その理由は，学生の修学状況を示す項目が多岐に渡ること，また，それらの項目の関係性が一意ではなく，どの項目がどの結果の要因と成り得るかが単純には判断できないことが挙げられる．例えば，「１年次の GPA と４年間 GPA には強い正の相関がある」，「音楽系サークルに所属している学生は GPA ランクが２ランク程度低い」，「就職講座の出席率の高い者は低い者に比べて内定率が 20 ポイント高い」など，２次元（項目），３次元程度の関連性は論じることができてもそれ以上の次元での関連性は複雑で調べようがない．ましてや EM に必要な退学者予測モデルなど立てようがなく，これら集めた学修データやそれに関連するデータはほとんど意味のないものとなっていた．

２．研究の目的

上記課題を解決するための本研究の目的として次の４つを想定した．

（１）最初の目的として，IR のみならず，機械学習の教師ありデータとして処理するために，学生の学修データをどのような形式，構造で構成することが最適かを検証する．具体的には，申請者が関わった様々なデータの形式や構造を分類し，これらを現場レベルでクレンジングするための方法を確立しマニュアル化して，まずは所属大学において共有化し徹底を図る．IR に取り組む大学でもこれらクレンジング作業は担当者の個々のスキルに依存し，経験則のような暗黙知的側面が強い．これを形式知（マニュアル）化することは学術的にも意義があり独自性も有するものと判断される．

（２）2 番目の目的については，AI などの機械学習や分析ツールをクラウド上でパッケージとして提供している商用クラウドサービスが多数登場しており，これら商用のクラウドサービスを利用して機械学習が可能かどうか検証する．合わせて，Python などのスクリプト言語でも同じ結果が得られるか比較検証する．

（３）3 番目の目的については，商用クラウドサービスの機械学習の有効性を検証した後に，それら判断結果が本当に大学の教職員の持つ経験則やカン，いわゆる暗黙知と合致するのか否かの検証を行う．過去の 8 千件のデータを教師ありデータとして機械学習させ，今後の入学予定者 8 百件/年の予測データを随時入力し，例えば退学予測，休学予測，GPA 上昇，低下などいくつかの項目でどのような予測がなされるかを出力し，これらを検証時点での教職員の予測，6 か月後，1 年後，2 年後時点での実際値と比較し，機械学習，教職員，実際値の比較検証を行う．

（４）4 番目の目的については，EM システムの構築を行い，何か予測したい情報がある場合には，作成した機械学習のモデルをインターネット経由で呼び出し，予測や判断結果を得ることが可能となるよう Web API(Application Programming Interface)を構築し実装する．それによって，各大学は自前のシステムで IR もしくは EM 機能だけを設け，機械学習や予測機能はクラウドサービスを利用するなど，大学の状況に合わせてシステムを柔軟に構築することが可能となる．

３．研究の方法

研究開始時から，学内における様々なデータを収集し，集めたデータを全て学籍番号を主キーとして統合し，さらに機械学習等の分析を行うためにデータのクレンジングを行う．次にこれら統合化されたデータを，一つはクラウドサービスによる分析ツールをいくつか試し結果の可視化をすることによって，有用な知見が得られる手法を見出し，実例を積み重ね学内で共有化を行う．もう一つは，機械学習による分析のターゲットを学生の退学予測に定め，その予測を行うために最適な機械学習予測モデルを選択し，出力された結果を評価する．さらにはその機械学習による退学予測結果を今後の EM に用いるためのシステムを構築し，実際に実装しその効果を検証する．

４．研究成果

（１）学修等データのクレンジングとクラウドサービス分析ツールによる可視化と共有

データのクレンジングは学籍番号を主キーとし，また，学内の様々な部署で保有しているデータを収集する仕組みを構築し，定期的にデータを集約する方法を確立した．

また，クラウドサービス上の分析ツールについては，Tableau と Microsoft PowerBI を導入し，両者を比較検討した結果，学内の多くのユーザがライセンス的に安価に利用しやすく，また，MicrosoftOFFICE 製品と親和性が高い Microsoft PowerBI に統一して分析を行った．

PowerBI は，Web 上で利用者が動的に分析指標を任意に選択することが可能で，学内の複数の部署で IR 活用ツールとして利用が開始された（図１）．

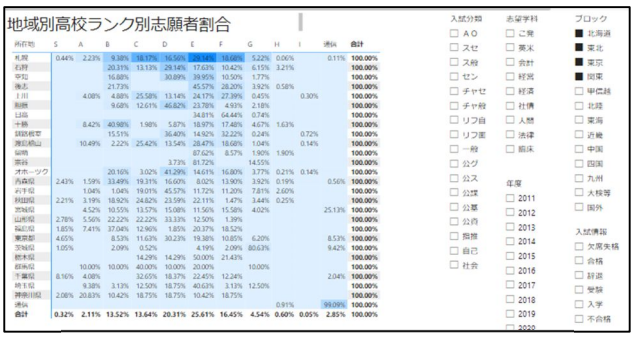

図１　PowerBI による分析結果の例

（２）機械学習による分析と退学予測ツールへの実装

以下の研究成果は，ほぼ，引用文献[1]で発表した論文で成果をまとめており，より詳細な内容は引用文献[1]を一読いただきたい．

①機械学習のためのデータ変換

機械学習においては，一連のサンプルに基づいて入力データを既知の値に写像する為の規則を学習することで予測を行う“教師あり学習”，目的値を用いずに入力データの重要な変換を見つけ出す“教師なし学習”，エージェントがその環境に関する情報を受け取り，なんらかの報酬が最大になるような行動の選び方を学習する“強化学習”など，様々な種類が存在するが[1]，本研究では学生の退学確率を予測するという目的に準えて，教師あり学習を採用した．

この教師あり学習に使用する変数（目的変数，説明変数）については，目的変数は，学生の在籍状況であり，卒業するか，退学するかの二値をとるように加工した．実際の本システムの活用場面においては，１年次終了の際に実施される担任教員と学生との個別面談での指導支援を見据えていることから，説明変数には高校までの情報（出身高校の偏差値，入試方式など）と入学から１年次終了時までの情報（１年次の GPA など）を用いることとした．一方，２年次以降の情報（２年次の取得単位数，就職先の業種など）は仮に収集可能であっても，予測に用いることは行わないこととした．

モデルの構築には，2011 年度から 2017 年度に入学した学生の情報を用いた．なお，１年次が終了する以前に退学した学生については，本実践の目的である，１ 年次終了時の指導に活かす，ということが不可能である為，除外した．

②機械学習による退学予測モデル（アルゴリズム）の選択

機械学習を用いて学生の退学確率を高い精度で予測する為，複数の予測モデルをまずは構築し，その比較を行った．

分類問題のための機械学習アルゴリズムとしては様々なものが提案されているが，本研究では，ロジスティック回帰（Logistic Regression），サポートベクターマシーン（Support Vector Machine，以下 SVM），ランダムフォレスト（Random Forest）に加え，勾配ブースティング（Gradient Boosting Machine）の代表的な実装である LightGBM（Light Gradient Boosting Machine）と XGBoost（eXtreme Gradient Boosting）の５つを検討した．

これら複数の予測モデルを比較する上で，共通の評価指標が必要であるが，退学者に比べて卒業者の圧倒的に多い不均衡データの場合，仮に全ての学生を「卒業する」と予測しても正解率は 86%ほどになり，評価としては適切ではない．従って，AUC（Area Under the ROC Curve）を利用した．AUC は０から１までの値を取り，１に近いほど判別能が高いことを示す．算出には，予測値を正例と判断する閾値を１から０に動かす中で，偽陽性率（負例を誤って正例と予測する割合）と真陽性率（正例を正しく正例と予測する割合）の関係を図示する ROC 曲線（Receiver Operating Characteristic Curve）における，曲線の下部の面積を用いる．

ランダムフォレストであれば，構築する決定木の個数や深さの最大値など，各アルゴリズムには固有のハイパーパラメータが存在し，予測精度を高める上でこれらを適切にチューニングすることが求められる．また，機械学習の評価においては，学習に用いていない未知のデータに対する予測性能である“汎化性能”を比較する必要がある[2]．ハイパーパラメータのチューニングと学習済みモデルの汎化性能の評価を同時に行う方法として，入れ子構造の交差検証（Nested Cross-Validation）を用いた．

以上の検証によって得られた予測精度は表１のようになった．これから最も精度が良かった，LightGBM を以後では採用することとした．

表１　アルゴリズム毎の AUC

アルゴリズム	AUC
ロジスティック回帰	0.865
SVM	0.848
Random Forest	0.863
LightGBM	0.870
XGBoost	0.869

③機械学習モデルの評価

予測モデルに Light GBM を用いる場合には，特徴量重要度（Feature Importance）を算出することが可能となる．これを用いることにより，どのような変数が退学の予測に寄与しているのか，確認することが可能である．特徴量重要度にはいくつかの算出方法があるが，ジニ不純度（Gini Importance）を基にした方法では，各変数の値を閾値として決定木の分割を定義した際の，ジニ不純度の減少度合いによって，変数の重要度を定義している．

この方法で得られた特徴量重要度(図 2 参照）によると，1 年次で取得した単位数，1 年次の GPA の順で，予測への寄与度が大きいことが読み取れる．この 2 変数が退学確率に影響することは，“暗黙知”とも一致する他，卒業者と退学者の評定分布を比較しても明らかである（図 3，図 4 参照）

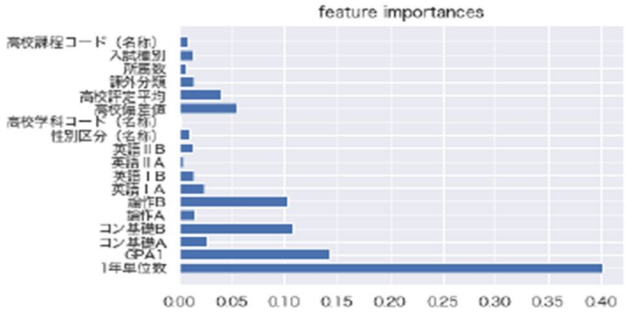

図 2　特徴量重要度

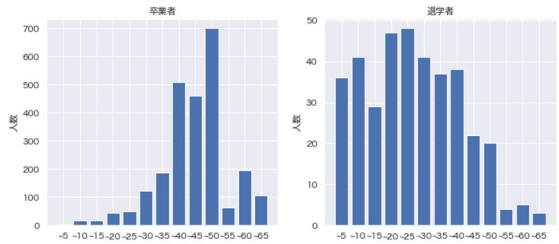

図 3　卒業者・退学者別の 1 年次取得単位数

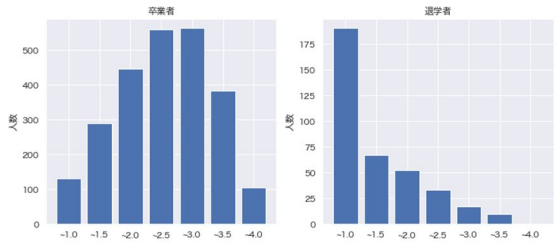

図 4　卒業者・退学者別の 1 年次 GPA

また，他の科目からの重要度を検証すると，コンピュータ基礎 B，論述作文 B の評定が重要視されていることがわかる．両者とも 1 年次後期に開講される講義であり，1 年次終了時直近の学生の状態を表していることが，前期に開講されるコンピュータ基礎 A，論述作文 A に比べて重要度が高く出ている要因だと考えられる．また，他の必修科目と比較しても，この 2 科目は不可率が高い傾向にあり（表 2 参照），これらの単位の取得有無が退学と卒業を決定付ける重要な分かれ道になっている．

表 2　1 年次科目と不可率

科目	不可率
コンピュータ基礎 A	11.7%
コンピュータ基礎 B	19.0%
論述作文 A	6.11%
論述作文 B	15.7%
英語ⅠA	9.29%
英語ⅠB	7.34%
英語ⅡA	11.8%
英語ⅡB	10.4%

④退学予測ツールの実装

構築した予測モデルを，実際の学生のサポートに使用する EM システムとすべく，退学予測ツールのプロトタイプを製作した．本システムは，Microsoft Excel で作成されたユーザーインタフェースに加え，Visual Basic for Applications (VBA)を用いて記述されたソースファイル，Python を用いて記述されたソースファイル，及び Light GBM における学習済みモデルが pickle 形式で格納されたソースファイルから構成される．

例として，真面目な学生，不真面目な学生をイメージした学生情報を入力し，予測を行った場合の結果を図 5 に示す．それぞれ退学可能性が 70%，12%と出力され，直感にも則した退学確率が表示されていると見られる．このような仕様を実現することで，利用を想定する教職員においては，裏側で機械学習による処理が行われていることを意識せずとも，必要な情報を入力すれば結果を得ることが出来る．すなわち，コーディングや機械学習に関する知識や技術が伴わない場合でも，問題なく利用可能ということになる．また，学習済みモデルを使用しており，裏側では学生情報に基づく推論のみが行われているので，完了ボタンをクリックしてから結果が出力されるまでの工程は，1 秒足らずで完了する．

図 5　予測ツールの画面例

⑤退学予測ツールの評価

本来であれば，試作したプロトタイプを実際の修学指導現場で使用し，教職員のカンとの比較を行い，本システムの評価をすべきであったが，学内的な実証環境が整わず，本研究期間内には

検証できなかった．しかしながら，2011 年度から 2017 年度までのデータを訓練データとテストデータに分割して教師データとして使用していることから，2018 年度以降のデータがあれば純粋に本システムの予測の評価が可能となるのでその一例を紹介する．

特に現場での適用を見据えて，より解釈性の高い混同行列を用いた評価，考察も行った．混同行列では，A.本システムで退学と予測して実際に退学した学生の数（真陽性，True Positive），B.退学と予測したが実際には退学しなかった学生の数（偽陽性，False Positive），C.卒業と予測したが実際には退学した学生の数（偽陰性，False Negative），D.卒業と予測して実際に卒業した学生の数（真陰性，True Negative），を 2×2 のクロス表で提示するものである．

2022 年 3 月に卒業年であった 2018 年度に入学した学生(前処理後，540 人）の退学確率について，閾値を 0.1 から 0.5 と変更して得られた混同行列を示す（表 3）．今回は「退学確率」を最終結果としており，何%以上を退学と判断し指導対象にするのか，何%未満を卒業と判断するのか，閾値を変更してシミュレーションを行うことが出来る．例えば，退学確率 80%を閾値とすると，本当に退学しそうな学生だけを指導することになり，B が小さくなって指導の負荷は下がるが，逆に C が大きくなり見落としが増える．閾値を低くすると，少しでも退学可能性がある学生には指導をするので，見落としは減る分，指導の負荷が増える，というようなトレードオフが生じることがわかったが，現場の状況に応じて閾値を調整し対象者を任意に選択できるようになった．

表 3　閾値と混同行列

閾値	実際の結果	退学と予測	卒業と予測	指導対象者数	見落とし数
0.1	退学	71	12	172	12
	卒業	101	356		
0.2	退学	56	27	92	27
	卒業	36	421		
0.3	退学	45	38	57	38
	卒業	12	445		
0.4	退学	39	44	42	44
	卒業	3	454		
0.5	退学	30	53	30	53
	卒業	0	457		

この結果を見ると，閾値 0.1 に設定した場合，指導対象者数は 172 名（全体の約 32%）と多くはなるが，卒業と予測したのに退学した学生，すなわち見落とし数は 12 と少なくなった．また，閾値 0.5 の場合は，指導対象者数 30 名（全体の 5%強）と少なく，退学と予測し実際に退学した学生も 30 名であるが，逆に見落とし数は 53 名に達しており，これでは指導の意味がない．現在行われている指導対象者の範囲は学科によっては対象者の 5 割以上のところもあるので，閾値 0.1 でも実際の運用は可能と考えられる．さらに言えば，退学可能性 20%～50%の学生にはメールを送る，50%～70%の学生には面談を行う，70%以上の学生には保護者にも働きかける，といった閾値をコントロールして段階的な処置を行うといったことも実践上は可能である．

＜引用文献＞

[1] 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トランザクションデジタルプラクティス，Vol.4 No.2 2023

[2] Francois Chollet :Python と Keras によるディープラーニング，株式会社マイナビ出版，2018

[3] 山口達輝,松田洋之:機械学習＆ディープラーニングの仕組みと技術がしっかりわかる教科書，技術評論社，2019

５．主な発表論文等

〔雑誌論文〕　計2件（うち査読付論文　1件／うち国際共著　0件／うちオープンアクセス　1件）

１．著者名 石川千温	４．巻 33
２．論文標題 AIの身近に迫る問題	５．発行年 2021年
３．雑誌名 日本商業教育学会北海道部会報	６．最初と最後の頁 2,3
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 無
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 石川千温，石本翔真	４．巻 4
２．論文標題 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	５．発行年 2023年
３．雑誌名 情報処理学会論文誌デジタルプラクティス	６．最初と最後の頁 1-8
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

〔学会発表〕　計4件（うち招待講演　2件／うち国際学会　0件）

１．発表者名 石川千温
２．発表標題 札幌学院大学におけるIR活動の事例
３．学会等名 北海道科学大学IR研修会（招待講演）
４．発表年 2022年

１．発表者名 石川千温
２．発表標題 経営学科のIRを考える
３．学会等名 札幌学院大学総合研究所経営部会
４．発表年 2021年

１．発表者名
石川千温

２．発表標題
AIの身近に迫る問題

３．学会等名
日本商業学会北海道部会

４．発表年
2021年

１．発表者名
石川千温

２．発表標題
札幌学院大学のIRの取り組み

３．学会等名
２０１９年度IDE大学セミナー（招待講演）

４．発表年
2019年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

2019-20年度札幌学院大学IR報告書
https://www.sgu.ac.jp/information/j09tjo00000fhvq5-att/j09tjo00000fhvs3.pdf

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関