機械学習と SHAP を用いた教育に関するアンケートの分析

山田 航大 [1)], 小林 幹 [1)], 青木 和昭 [2)]

1) 立正大学経済学部
2) 立正大学地球環境科学部

miki@ris.ac.jp

Analysis of a questionnaire on education using machine learning and SHAP

Kota Yamada [1)], Miki Kobayashi [1)], Kazuaki Aoki [2)]

1) Faculty of Economics, Rissho Univ.
2) Faculty of Geo-Environmental Science, Rissho Univ.

概要

本研究では学生の成績(GPA)に影響を与えると考えられる自身の学習行動だけでなく、社会的、遺伝的な要因など多面的な要因について機械学習を用いて分析し、予測に大きく寄与する要因を把握するためにモデル解釈ライブラリ(SHAP)を用いることによって特徴量間の関係を公平に判断する。海外の事例を用いた本研究では学校の欠席数や過去の落第回数が成績に大きく寄与するという直感に合う結果が示された。さらに、本手法を用いて、他の要因がいかに成績と関連するのかについて議論を行った。

1 はじめに

近年、働き方やデジタル化を考慮した改革が行われるようになり社会のあり方が変化している。これは大学においても同様でありオンライン学習や e-ラーニングの導入などによって日々新たな学習方法が検討、導入されている。オンライン学習や e-ラーニングの導入により、学生の学習データを詳細に獲得できるようになり、それらを用いてどのような要因が学生の成績に影響を与えるのかを把握することが比較的容易に可能となった。これらのことは関連研究においても学生の学力に影響を及ぼす学習要因などとして日々研究されている。しかしながら、関連研究において、さまざまな要因による影響やそれぞれの相互作用を考慮した分析が不十分であると考えられる。そこで本研究では、学生の遺伝的な要素や、社会的な要素など多面的な要因を考慮しながら、機械学習を用いて学生の成績を予測する。さらに、用いたモデルに対して学生の成績に影響を与える特徴量を把握することを行う。これにより、学生の学習効率の向上を図ることができると、同時に今後の授業方針を決定する上での参考になると考えられる。

2 先行研究

学生に対して成績と勉強方法に関するアンケートや分析は多くの大学で行われている。心理学の観点を重視したものや統計学を用いた調査など数多く行われておりその内容は多岐にわたっている。藤田保健衛生大学医学部における入学後の成績に影響を与える要因は何かという分析において入学後の成績が入学直後の基礎学力のみに影響されるのではないこと、および 1 年次での学習態度、意欲がその後の成績を左右する点で極めて重要であると分析されている[1]。平により、相関分析を用いて学生の学習経験に起因するとする学習経験要因と学生自身の特性や学びの心的傾向性といった学生内要因が成績に影響を与え、潜在成長曲線モデルを用いた分析では特に高校までの統合的学習経験や多様な理解を学習で重視するような生徒が良い成績を獲得することが報告された[2]。

3 モデルについて

3.1 使用するモデル

本研究で用いる学習モデルを決定するために、サポートベクターマシン(SVM)、重回帰分析(MRA)、ランダムフォレスト(RF)、勾配ブースティング決定木(GBDT)の 4 種類を 4 章で用いるデータを使用して比較した。学習データとテストデータの比率をいくつか変更し、その予測精度を比較した結果を表 1 に示す。

表 1　各モデルの学習データとテストデータの割合（1 行目）を変化させた場合の予測精度。標柱の数値は予測値(予測された数学の成績)と正解値(実際の数学の成績)との差の絶対値平均で計測している。値が小さいほど予測精度が高い。成績は 0 から 20 までの評価である。

	6:4	7:3	8:2	9:1	Ave.
SVM	3.356	3.386	3.496	3.982	3.555
MRA	3.448	3.522	3.456	3.814	3.560
RF	3.080	3.101	3.053	3.260	3.124
GBDT	3.591	3.465	3.695	4.047	3.700

上記の表より、RF の予測精度が高く、他 3 種類は同程度の精度であった。本研究では予測精度、およびモデルの解釈しやすさという観点から、RF および GBDT をモデルとして用いることとした。データについては、学習データ 7 割、テストデータ 3 割とした。これらの木構造を持ったモデルはアンサンブル手法を用いており、過学習を抑制しながら高い汎化性を持っていることも、RF と GBDT を選定した重要な理由である。

3.2　ランダムフォレスト[5]について

ランダムフォレストとは、決定木とアンサンブル学習の 2 つを組み合わせアルゴリズムである。1 つ目の決定木とは予測、分類などにおいてデータに対し yes/no(数値なら大きい/小さいなどを考慮するときもある)で質問を階層的に繋げていき、その質問に対して順番に答えていくことで最終的な答えにたどり着く仕組み、構造を持っているもので、この階層構造が木の枝が分岐しているように見えることから決定木と呼ばれている。

2 つ目のアンサンブル学習とはより良い予測精度を得るために複数の学習アルゴリズムを組み合わせる技術である。分類問題の場合には複数の学習器の多数決によって予測結果を決め、回帰問題においては複数の学習器の平均を取ることで予測を行う。これら 2 つのアルゴリズムを用いて複数の決定木を集めて、アンサンブル学習によって予測を行うような手法をランダムフォレストという。

3.3　勾配ブースティング決定木[6]について

勾配ブースティング決定木は勾配、ブースティング、決定木の 3 つによってこの手法を説明することができる。決定木はランダムフォレスト(3.2)において説明したとおりである。勾配とは勾配降下法を用いたものである。一般的に、機械学習アルゴリズムの目的は正確な予測を行うことであるが、予測が正確かどうか判断する基準となるのは誤差の大きさであり、この誤差を小さくする方法の 1 つが勾配降下法である。この方法は、損失関数上の勾配(傾き)が最小化となるものを求めていくことでパラメータ最適を行う手法である。勾配は変数の増加に対して関数が増減する方向と量を示している。

そして 2 つ目にブースティングというアンサンブルの手法を用いている。このアルゴリズムは複数のモデル使い直列的にそれらを繋げる手法を取る。モデルを直列的に繋げ、前に作ったモデルの結果を参考に次のモデルを構築することでより良い精度を持った学習器を生成することができる。以上、3 つの概念を組み合わせた手法である。

3.4　SHAP[7]について

SHAP(SHapley Additive exPlanations)はゲーム理論のシャープレイ値を機械学習モデル説明に応用し、予測値に対する説明変数の寄与度を計算する手法である。SHAP を用いることで学習で得られた特徴量の重要度の把握、特徴量間の関係などを得ることができる。

4　結果

4.1　使用したデータ

海外におけるアンケート結果について SHAP を用いて分析する。ここで用いたデータは、UC Irvine Machine Learning Repository[4]というカリフォルニア大学アーバイン校によって機械学習用のデータセットとして寄付されているもののうち、Paulo Cortez 氏によって調査された Student Performance というポルトガルの 2 つの中学校の学生の数学の成績とその他の要因に関するアンケートであり、生徒の数学の成績、社会的特徴に関する質問など多面的な質問から構成されている。アンケート結果から数学の成績を予測する機械学習を行う。395 人分のデータを学習フェーズ 7 割とテストフェーズ 3 割に分けて使用する。使用するモデルは勾配ブースティング決定木とランダムフォレストを使用する。(3 章 1 節を参照)選定対象となるモデルの中で精度の良かったモデルに SHAP を適用し、モデル解釈を行う。また、SHAP は学習フェーズとテストフェーズのどちらに対しても使用可能であり、学習フェーズに適用した場合、モデルがどの部分に注意を払って学習しているか、どの特徴量が予測に最も寄与していのか把握することができる。そして、テストフェーズに適用した場合は、モデルが未知のデータに対してどの特徴量を重視して予測しているかを把握することができる。本研究においては、学生の成績に影響が大きい要因を把握したいため学習データに

SHAP を適用して解析を行う。

4.2　木構造について

今回、使用する GBDT とランダムフォレストは、どちらも木構造を持っているため、この節で木構造の掲示を行う。勾配ブースティング決定木やランダムフォレストのような木構造を持ったモデルは 3.2 と 3.3 において示したように以下の図のような木を複数持ち、この末端の結果を用いて多数決を取る形で結果を出力している(この末端を葉ノードなどと呼ぶ)。

図 1　本研究で用いた GBDT の 1 番目の木構造

4.4　shap 値の計算

この節では勾配ブースティング決定木とランダムフォレストにそれぞれ SHAP を適用して値の計算を行う。表 2 は用いられる特徴量に関する詳細情報である。

図 2 は、勾配ブースティング決定木に SHAP を適用した計算結果から特徴量の影響をグラフ化したものである。図の縦軸に特徴量名が書いてあり上からその特徴量が予測に与える影響の大きい順(貢献度が高い)に並んでいる。横軸には shap 値のプラス(予測に対して正の影響を与える),マイナス(予測に対して負の影響を与える)を表している。赤い点、青い点、紫の点はそれぞれ特徴量の実際の一人一人の回答である。わかりやすい例として下位 6 番目の特徴量(表 2 参照)である romantic(恋人がいる(1)、いない(0))を見てみると赤い点(特徴量の値が大きい)では shap 値はマイナスの値をとっている。青い点(特徴量の値が小さい)では shap 値はプラスの値をとっている。このようなことから恋人がいると成績に負の影響を与えやすいということがわかる。また、点が積み上げられているところでは密度が高いことを表している。このグラフの上位に来ている特徴量である absences(学校の欠席数)は 4 章 2 節の図 1 で示した木構造においても上位に来ている。このことから SHAP の結果が妥当であることが示唆される。

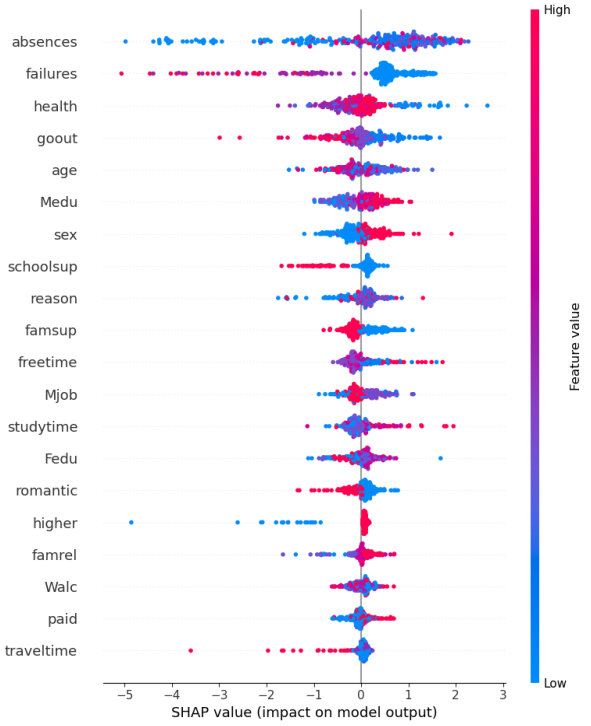

図 2　shap 値の計算結果(勾配ブースティング決定木)

続いてランダムフォレストに SHAP を適用した時の結果を示す。モデル自体は勾配ブースティング決定木と同様に木構造を持っているため shap 値の計算結果によって上位に登場する特徴量はあまり変化がないと考えられる。また次の節で shap 値の計算による特徴量エンジニアリングの変化を見るためにもこの節でランダムフォレストに SHAP を適用した計算結果を表示する。図 2 と図 3 より木構造のモデル間では shap 値の計算による特徴量の変化はほとんど見られなかった。どちらのモデルも absences(欠席数)、failures(落第数)は上位に来ており一貫して重要だということが示された。また下位の特徴量に行くと分散が小さくなりそれほど成績に対して影響が少ないことがわかる。

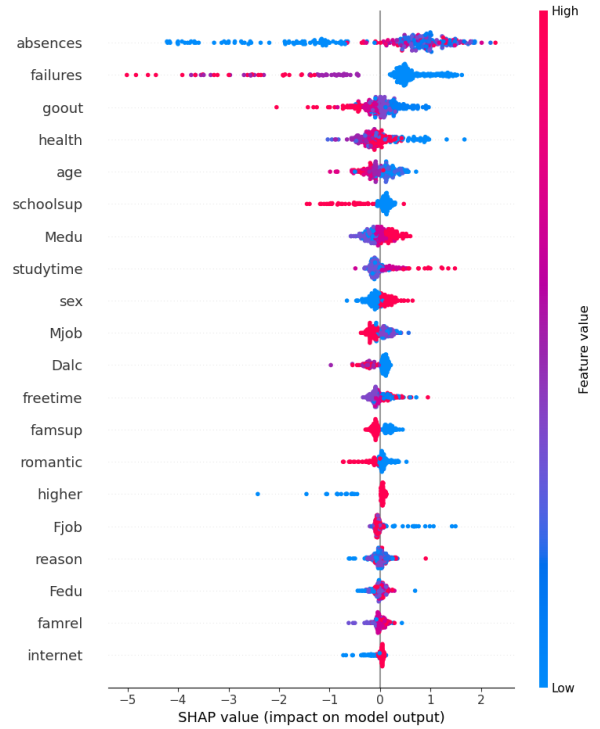

図 3　shap 値の計算結果(ランダムフォレスト)

absences	学校の欠席数(0-93)
failures	過去の落第回数(0-4 以上も 4)
health	現在の健康状態(1:非常に悪い-4:非常に良い)
goout	友達と遊びに行く頻度(1:非常に少ない-4:非常に多い)
age	年齢(15-22 まで)
medu	母親の教育(0:なし、1:初等教育(4 年生、2:5 年生から 9 年生、3:中等教育、4:高等教育)
schoolsup	追加教育サポート(1:yes、0:no)
Walc	週末のアルコール摂取量(1:非常に少ない-5:非常に多い)
reason	学校選択の理由(5 家から近い、6:学校の評判、7:コースの好み、10:その他
famsup	家族教育サポート(1:yes、0:no)
freetime	学校後の自由時間(1:非常に少ない-5:非常に多い)
Mjob	母親の仕事(5:教師、6:医療関係、7:公務員、8:家庭、10:その他)
studytime	週あたりの学習時間(1:2 時間未満、2:2-5 時間、3: 5 -10 時間、4:10 時間以上)
addres	住所タイプ(1:都市、0:田舎)
romantic	恋愛関係(1:yes、0:no)
higher	高等教育を受ける意向(1:yes、0:no)
famrel	家族との関係(1:非常に悪い-5 非常に良い
Internet	家庭でインタネットが使えるか(1:yes、0:no)
paid	追加の有料クラス(1:yes、0:no)
traveltime	家から学校までの通学時間(1:15 分未満、2:15-30 分未満、3:30 分-1 時間、4:1 時間以上)

表 2 本研究の結果に登場した各特徴量について

5.3　特徴量エンジニアリングについて

特徴量エンジニアリングを行うことで予測精度の変化と見られると同時に関連性が高い特徴量のみを使用することで理解のしやすさが向上することができる。本研究では shap 値の結果の上位から順に追加していく場合と shap 値の結果上位から一つずつ適用した時による変化を検証する。まず初めに shap 値の上位から順に説明変数に追加していった場合の予測精度をグラフ化していく。このグラフを見ることで特徴量同士の相互関係などを確認することができる。以下の図が予測精度を表したものであるが shap 値の上位から増やしても予測精度はあまり変化しないことが確認できた。

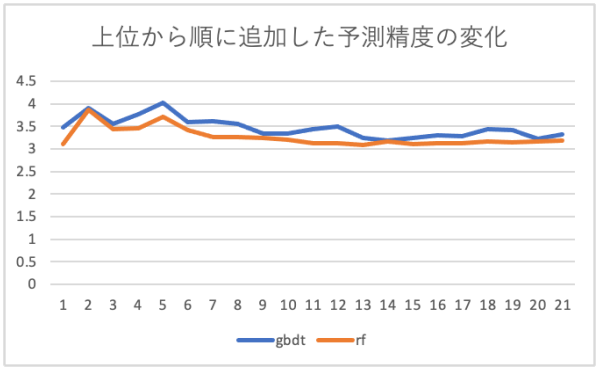

図 4　shap 値の上位から順に一つずつ説明変数に増やしていったときの予測精度の変化

続いて、shap 値の上位から順に一つずつ説明変数

に適用した場合の予測精度の変化を見ることで重要な特徴量が確認することができると同時にノイズとなるような特徴量があった場合最適な特徴量を選択し、モデル学習を行うことができると考えられる。図５が予測精度を表したものであるが図3で示したshap値の上位から順に説明変数に適用した場合よりも予測精度に変化はなかった。

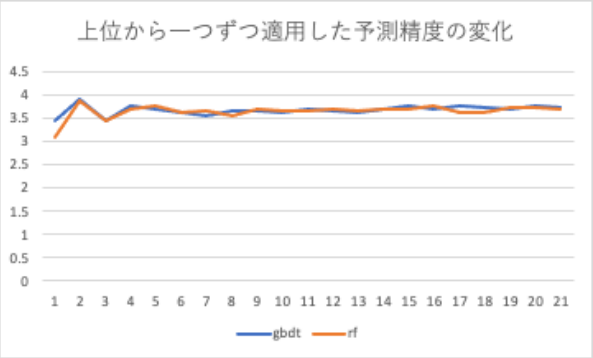

図 5 shap 値の上位から一つずつ説明変数に追加したときの予測精度の変化

以上より shap 値の結果から予測精度の向上は見られなかった。また、特徴量間同士でも予測精度を向上させる作用などは見られなかった。SHAPの結果の上位に来ていた特徴量を適用しても予測精度に変化は見られずノイズとなるような特徴量の特定も行えなかった。

6　考察と今後の展望について

本研究では学生の成績に寄与する要因を多面的な質問から分析し、それぞれが成績に対してどのように作用寄与するのか検討を行った。その結果、以下の点が示されたといえる。一つ目は自身の行動に関することが成績の変化に大きく寄与していると考えられる。先行研究では、学生自身の行動によって成績に影響があることが示唆されていたが、本研究では両親の職業などという経済的な要因や両親の学歴という遺伝的な要素が少なからず成績に影響していること、それ以上に影響が強いものとしてSHAPの結果から欠席数や過去の落第数、友達と遊びに行く回数や自身の健康状態という要因が成績に強く影響を及ぼしていることがわかった。このような要因が追加の有料クラスを受けているかという paid などの項目よりも成績への影響が大きいため有料で追加のクラスを受けるよりもまず学校への欠席数を減らし現在受けている授業に参加することが大切であると考えられる。二つ目は成績に影響を与える大きさが小さい要因は、学生間における shap 値の分散が小さいため、教育に与える影響が軽微だと言える。具体例として図3の下位1番目であるinternetが挙げられる。近年 e-learning などといったインタネット環境が必要になってくるような学習方法が学校教育でも導入されているが本研究の結果から一概には言えないがインターネットを利用できる環境があるかないかは、成績に与える影響は少ないものだと言える。このようなことから影響が少ないような要因について取捨選択を行い教員の仕事や学生の成績向上の効率化を図ることができると考える。今回は海外の事例を用いたが今後、立正大学の学生を対象とした調査を行い日本の大学生の分析を行なっていく。

参考文献

[1] 中島昭　長田明子　石原慎　大槻眞嗣　橋本修二　小野雄一郎　野村隆英　松井俊和、入学後の成績に影響を与える要因は何か、医学教育、39巻、6号、p397-406、2008年

[2] 平知宏、入学後成績推移における学習経験要因と学生内要因の影響、大学入試研究ジャーナル、32巻、p278-285

[3] Scott Lundberg Su-In Lee 、A Unified Approach to interpreting Model Predictions、NIPS 2017,Advances in Neural Information Processing Systems30,2017

[4] Paulo Cortez、Student Performance、UC Irvine Machine Learning Repository、11/26/2014、https://archive.ics.uci.edu/dataset/320/student+performance

[5] L. Breiman, Random Forests, Machine Learning, Vol. 45, No. 1, pp. 5-32, 2001.

[6] A. Natekin, A. Knoll, Gradient boosting machines, a tutorial, Front. Neurobot., Vol. 7, pp. 1-21, 2013.

[7] L. Scott and S. Lee, A unified approach to interpreting model predictions, Advances in Neural Information Processing Systems, pp. 4765-4774, 2017.