N-025

学習者の受講データを使用したLightGBMによるドロップアウト予測
Analysis of Dropout Causes in Student History with Light GBM

宮田 大[1)]　大枝 真一[2)]
Dai MIYATA　Shinichi OEDA

1　はじめに

近年，学校やオンライン講義の学習履歴や行動履歴を収集するシステムが増えている．特に，無償で利用できる大規模公開オンライン講座（MOOC）[1]は，世界中でリリースされ，広く導入されている．また，MOOC は大学レベルの高度な知識を学ぶことが可能である．日本においても既に導入が始まっており，人材育成に力をいれている企業も少なくない．MOOC 全体の学習者数は，中国や一部地域を除いて，2021 年時点で約 2 億 2000 万人の学習者がいる．その反面，MOOC では手軽に受講可能であるために学習者はとても多いが，講座の途中でドロップアウトする利用者が多く存在し，その平均修了率は 7%[2] と非常に低いことが問題である．この問題に対し，MOOC は学習者の受講履歴や行動履歴をデータとして蓄積できるため，この問題を技術的に解決することが可能である．これは，Learning Analytics（LA）や Educational Data Mining（EDM）といった学習者の環境の改善をはかる研究が行われている．一般的に，MOOC は全てオンラインでの講義かつ無償であることが，ドロップアウト増加に起因していると考えられている．しかし，そういった要因を変更することは容易ではない．

そこで，本研究では，学習者の行動履歴から Light GBM でドロップアウト予測を行い，どの特徴量がドロップアウトに起因しているか分析を行う．また，その特徴量の重要度が正しいか検証する．

2　関連研究

2.1　CLSA

CLSA: A novel deep learning model for MOOC dropout prediction では，学習者の行動データを利用してドロップアウトの予測を行っている．従来のモデルと比較してより高い精度と F1 スコアを実現する CLSA と呼ばれるモデルを提案している．CLSA は，Convolutional Neural Network（CNN）を使用して特徴を抽出し，高次元ベクトルを生成する．生成した高次元ベクトルを Long Short-Term Memory（LSTM）を使って，時系列データを組み込み，静的なアテンション機構を用いて重みを算出していた [3].

1）　木更津工業高等専門学校 専攻科 制御・情報システム工学専攻 2 年
2）　木更津工業高等専門学校 情報工学科

2.2　MOOCVERSITY

MOOCVERSITY - Deep Learning Based Dropout Prediction in MOOCs over Weeks[4] では，MOOC における早期のドロップアウト問題を機械学習を用いて解決している．ディープラーニングアルゴリズムを使用して，学生データの週次履歴や学生の行動変化を考慮してドロップアウト予測をしていた．また，コースに対する評価が継続的に低下している学生に警告を行っていた．結果としては，既存のモデルと比べて高い精度で予測できていた．

2.3　LightGBM with Optuna

A Lightweight Method using LightGBM Model with Optuna in MOOCs Dropout Prediction[5] では，ユーザーの学習ログに対するデータマイニング技術を使用して予測を行い，ユーザーの行動からいくつかの有用な特徴を抽出し，Optuna チューニング手法を備えた Light GBM に基づく軽量な手法を提案した．

3　手法

3.1　決定木

本研究では，学生のドロップアウト予測だけでなくその要因分析を行うため，要因が視覚的にわかりやすい決定木を使用する．決定木 [6] とは，目的変数に影響を与える説明変数を明らかにし，説明変数を木構造として分析する手法のことである．木構造によって分類を行う為に人間にも視覚的に分かりやすい特徴がある．決定木の学習は，不純度を指標として条件分岐を作成する．この不純度の計算には，ジニ不純度やエントロピーが存在する．一般的なアルゴリズムの種類は以下の通りである．

- CART（Classification and Regression Tree）
- CHAID（Chi-squared Automatic Interaction Detection）
- C5.0

3.2　Light GBM

Light GBM（Light Gradient Boosting Machine）[7] は，決定木の勾配ブースティングの 1 種であり，中でも高速であることが特徴である．勾配ブースティング決定木は，複数の決定木を組み合わせるアンサンブル学習のうちのブースティングという手法を用いたものである．ブースティングは前の弱学習器の結果を次の学習データに反映させる．具体的には，各決定木のデータセットの誤差を最小化するように学習

する．各決定木の出力は，前の決定木の出力に誤差を加えることで得られる．Light GBM は，主に以下の仕組みを用いて高速を実現している．

- leaf wise（best first）
  損失が小さくなるようなノードから分割を行う．過学習しやすい問題があるが，途中で分割を止めることで解決している．
- histogram based
  値をヒストグラム化し，複数の値を一つの bin にしたものを分岐点とすることで高速化を実現している．
- Gradient-based One-Side Sampling（GOSS）
  勾配が大きいデータは全て使用し，勾配の少ないデータはランダムサンプリングをする．これによって，学習データの数を減らし，高速化を実現している．
- Exclusive Feature Bundling（EFB）
  相互に排他的な複数の特徴量を 1 つの bundle にまとめ，1 つの特徴量として処理を行うことで高速化を実現している．

4 データ概要

本研究では，KDD Cup 2015 で配布されたデータセット [8] を基にした Understanding Dropouts in MOOCs で使用されているデータセット [9] を使用する．本データは中国の大学の MOOC における 247 のコースを受講する約 15 万人の学生ログデータである．学習データは約 2900 万，テストデータは約 1300 万となっている．ただし，データは同一 enroll_id においても，アクション発生毎に記録されているため，前処理後の実際の学習データは約 16 万，テストデータは約 8 万となっている．データ構成は以下の通りである．

1. enroll_id
   username と course_id のペア
2. username
   ユーザの ID
3. course_id
   コースの ID
4. session_id
   セッションの ID
5. acttion
   14 種類のユーザアクティビティから構成されており，全てクリックやドラッグなどの動作を記録している． 具体的には，講義概要や講義の進捗，コメントなどコースに関するアクション，動画の再生や一時停止などの動画に関するアクション，課題に関するアクションがある．
6. object
   ユーザアクションに対応したオブジェクト ID
7. time
   ユーザアクションの発生時間
8. truth
   ユーザのドロップアウトラベルを示す．なお，1 がドロップアウト，0 が非ドロップアウトである．

5 実験内容

5.1 前処理

本データは同一の enroll_id においても，各アクションが発生するたび記録されている．そのため，同一の enroll_id を統一し，特徴量としてアクションの種類を追加する必要がある．また，新たに特徴量として連続するユーザアクションの時差を追加する．

5.2 手順

本研究では，MOOC Doropout Prediction Dataset を用いて，受講者のドロップアウトの予測を行う．実験手順は以下の通りである．なお，学習モデルのハイパーパラメータは optuna[10] を用いて，チューニングを行う．

1. Light GBM によるドロップアウトの予測.
2. 学習モデルにおける特徴重要度の算出.
3. 最も寄与率が高い特徴量を抜いた時の予測精度への影響の検証.

5.3 評価指標

出力された予測結果がテストデータにおいて正確に予測できるかを示す指標には，様々な手法が存在する．本研究の評価指標として，Accuracy，Precision，Recall，F-measure，AUC を使用する．

これらの評価指標は 2 値分類において，評価を区別するために表 1 の混同行列を使用する．

表 1: 混同行列

		実測値	
		Positive	Negative
予測値	Positive	TP	FP
	Negative	FN	TN

5.3.1 Accuracy

Accuracy とは，正解率をデータ数で割ったものである．値は 0 ~ 1 をとり，1 に近いほど良いモデルといえる．Accuracy は混同行列を用いると，以下の式 (1) で表せられる．

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{1}$$

5.3.2 Precision

Precision とは，予測で Positive である場合の正答している割合のことである．値は 0 ~ 1 をとり，1 に近いほど良いモデルといえる．FP を重視する場合に使用され，誤検知のしにくさを表す．Precision は以下の式 (2) で表わされる．

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

5.3.3 Recall

Recall とは，正解値が Positive であるデータのうち，予測が正解である割合のことである．値は 0 ~ 1 をとり，1 に近いほど良いモデルといえる．FN を重視する場合に使用され，再現率を表す．Recall は以下の式 (3) で表わされる．

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

5.3.4 F-measure

F-measure とは，Precision と Recall がトレードオフの関係であることから，2 つの値を調和平均した値である．値は 0 ~ 1 をとり，1 に近いほど良いモデルといえる．F-measure は以下の式 (4) で表わされる．

$$F - measure = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{4}$$

5.3.5 AUC

AUC とは ROC 曲線下の面積のことであり，ROC 曲線とは TPR（True Positive Rate）と FPR（False Positive Rate）によって描かれるグラフである．TPR とは正のデータを正と予測した割合であり，FPR は負のデータを正と誤予測した割合のことである．TPR と FPR はそれぞれ以下の式 (5)，式 (6) で表わされる．

$$TPR = \frac{TP}{TP + FN} \tag{5}$$

$$FPR = \frac{FP}{FP + TN} \tag{6}$$

5.4 特徴重要度

特徴量重要度とは，ある特徴量で分割する際にどのくらいジニ不純度を下げられるかに基づいて計算され，分類への寄与度を測る指標である．本研究では，生徒のドロップアウトの要因を特徴量重要度から分析する．本研究では，SHAP（SHapley Additive exPlanations）[11] を用いてモデルへの理解を深める．SHAP は，協力ゲーム理論の Shapely Value を機械学習に応用したものであり，feature importance と違い予測精度への寄与度を測ることができ，モデル全体の理解を深めることができる．

6 実験結果

各評価指標の値を表 2 に示す．また，ROC 曲線のグラフを図 1 に示す．

表 2: 各評価指標の値

Accuracy	Precision	Recall	F-measure	AUC
0.828	0.857	0.940	0.897	0.810

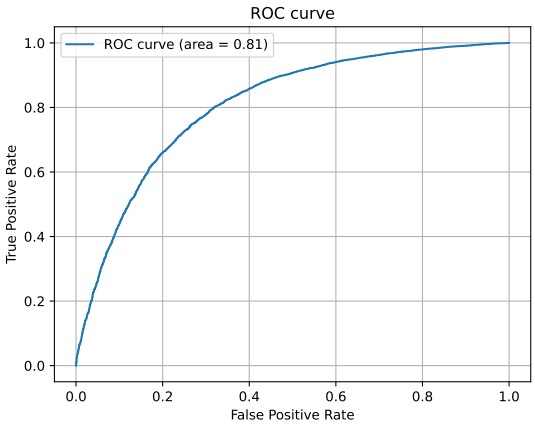

図 1: ROC 曲線

F-measure は 0.897 と比較的高い値となっているため，誤検知や再現性の面においても優れていることがわかる．また，AUC は 0.810 と高い値となっている．

本学習モデルの特徴重要度のグラフを図 2 に示す．

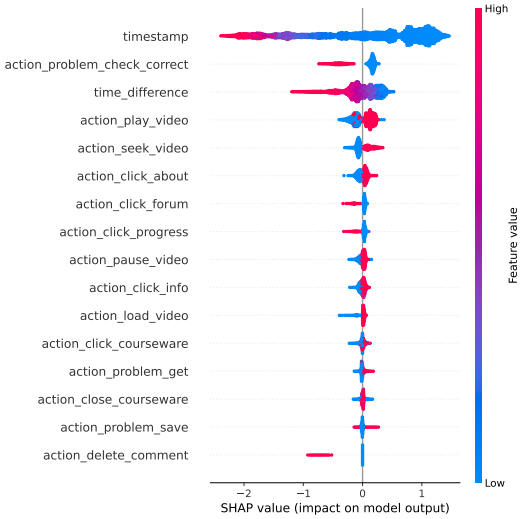

図 2: 説明変数ごとの SHAP 値

図 2 より，“timestamp” が最も予測結果への寄与率が高く，SHAP 値が大きい．また，“timestamp” は負の相関があることが分かる．次に “problem_check_correct”

と”time_difference” が寄与率が高く，“timestamp” と同様に負の相関がある．また，最も特徴重要度の高い “timestamp” を除いて実験した結果を表 3 に示す．また，その時の ROC 曲線を図 3 に示す．

表 3: timestamp を除いた各評価指標の値

Accuracy	Precision	Recall	F-measure	AUC
0.808	0.844	0.930	0.885	0.770

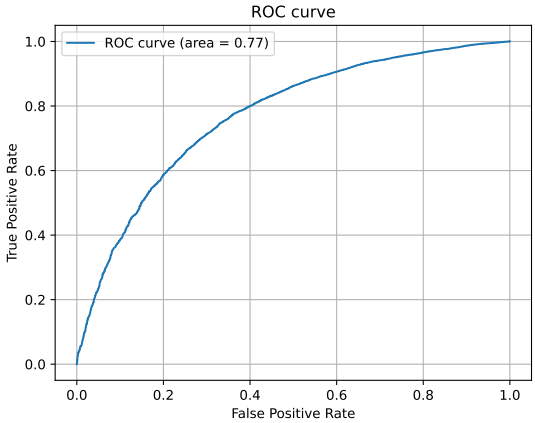

図 3: timestamp を除いた ROC 曲線

7 考察

学習済みモデルから特徴重要度を算出することで受講者のドロップアウトの原因を分析する．図 2 より，“timestamp” が最も特徴重要度が高く，次に “problem_check_correct” であった．“timestamp” は “action” と強い相関があることやコースの始まりや終わりなどに決まったアクションがあるため，特徴重要度が高いと考えられる．ドロップアウトする傾向にある学生は，課題に関してミスが多く，各アクション毎で時間を多く要している．また，表 2 と表 3 より，最も特徴重要度が高い “timestamp” を除いたが，AUC のみ 0.4 減少し，他の指標では精度にあまり影響が無かった．これは，Light GBM が複数の木を使用しているアンサンブル学習であることや “time_different” といった “timestamp” に関連した特徴量があるため，精度があまり落ちなかったと考えられる．

8 まとめ

本研究では，Light GBM を用いて受講者のドロップアウト予測とその要因を分析した．結果として，ドロップアウトしやすさは課題の正否やコースに要する時間が大きく要因していた．特徴重要度の算出には成功したが，これらの特徴量が真に原因しているかは断定できない．また，学生の行動は継続的に変化するため，このような予測を行う場合はより情報量の多い時系列データが望ましい．他にも，特徴量として学習者の性別などの個人属性の追加の検討が必要である．そこで，今後の課題として時系列データを用いた分析や Cohort Shapely[12] を用いた特徴量重要度の算出，MOOC 以外のデータやソースコードでの実験で検討している．

謝辞

本研究は JSPS 科研費 JP23K17604，JP24K00460 の助成を受けたものです．

参考文献

[1] MOOC.org．“Massive Open Online Courses”. https://www.mooc.org/，（2024-03-30）
[2] Christian G, Rocael HR, Chang V, Miguel M, “Attrition in MOOC: Lessons Learned from Drop-Out Students”, Proceedings of the international workshop on learning technology for education in cloud, 2014.
[3] Sinha T, Jermann P, Li N, Dillenbourg P, “Your click decides your fate: Inferring Information Processing and Attrition Behavior from MOOC Video Clickstream Interactions”, Proceedings of the EMNLP 2014 Workshop on Analysis of Large Scale Social Interaction in MOOCs, 2014.
[4] Vignesh Muthukumar, Dr. N. Bhalaji, “MOOCVERSITY - Deep Learning Based Dropout Prediction in MOOCs over Weeks”, Journal of Soft Computing Paradigm (JSCP) (2020)
[5] Kary Ng, Philip I S Lei, “A Lightweight Method using LightGBM Model with Optuna in MOOCs Dropout Prediction”, ICEMT ’22: Proceedings of the 6th International Conference on Education and Multimedia Technology
[6] J. R. Quinlan, “Induction of Decision Trees”, Machine Learning, Vol. 1, Issue 1, pp.81-106 (1986).
[7] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, Tie-Yan Liu, “LightGBM: A Highly Efficient Gradient Boosting Decision Tree”, Part of Advances in Neural Information Processing Systems 30 (NIPS 2017).
[8] Mooc Data． “KDD cup 2015”. http://moocdata.cn/challenges/kdd-cup-2015，（2023-06-30）.
[9] Wenzheng Feng, Jie Tang and Tracy Xiao Liu, “Understanding Dropouts in MOOCs”, AAAI 2019
[10] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, Masanori Koyama, “Optuna: A Next-generation Hyperparameter Optimization Framework”, KDD 2019 Applied Data Science track
[11] Scott Lundberg, Su-In Lee, “A Unified Approach to Interpreting Model Predictions”, NIPS 2017
[12] Masayoshi Mase（Research & Development Group, Hitachi, Ltd.）, “Explaining black box decisions by Shapley cohort refinement”, JSAI, 2022