Computer Science and Application 计算机科学与应用, 2019, 9(11), 2052-2064
Published Online November 2019 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2019.911231

Hans汉斯

Research on the Influencing Factors of User's Dropout in Theory and Practice Class

Mengjing Qi[1], Mingxian Wang[2], Lei La[1]

[1]School of Information Technology & Management, University of International Business and Economics, Beijing
[2]Office of Educational Administration, University of International Business and Economics, Beijing
Email: qmj1014@qq.com

Received: Oct. 30th, 2019; accepted: Nov. 12th, 2019; published: Nov. 19th, 2019

Abstract

Due to the high skipping rate and low completion rate of MOOCs, it is difficult for MOOCs to make full use of the advantages of Shared service economy to promote the reform of China's education model. Current scholars' prediction of skipping MOOCs has some problems, such as low accuracy and lack of targeted analysis of different courses. Select new feature variables and use XGBoost algorithm in integrated learning to model and predict the background data of different types of courses. Therefore, the factors influencing students' skipping classes of different types of courses are studied through comparison. The experimental results are compared with the traditional machine learning algorithm SVM, logistic regression and integrated learning algorithm AdaBoost. The research proves the effectiveness of XGBoost algorithm in the prediction of skipping MOOCs, and the difference of influencing factors of different types of courses. The results are of great practical significance to the improvement of platform retention rate and effective utilization rate of resources.

Keywords

Moocs, Prediction of Dropout, XGBoost

理论、实操慕课课程用户翘课影响因素研究

齐梦晶[1]，王铭娴[2]，喇　磊[1]

[1]对外经济贸易大学信息学院，北京
[2]对外经济贸易大学教务处，北京
Email: qmj1014@qq.com

收稿日期：2019年10月30日；录用日期：2019年11月12日；发布日期：2019年11月19日

文章引用: 齐梦晶, 王铭娴, 喇磊. 理论、实操慕课课程用户翘课影响因素研究[J]. 计算机科学与应用, 2019, 9(11): 2052-2064. DOI: 10.12677/csa.2019.911231

摘　要

慕课的高翘课率、低完成率使其难以发挥共享服务经济的优势以促进我国教育模式改革。当前学者对于慕课翘课的预测存在准确率较低和缺乏不同课程针对性分析等问题。选择新的特征变量，借助集成学习中的XGBoost算法对不同类型课程的后台数据进行建模和学员翘课预测。从而通过对比研究出不同类型课程的学员翘课影响因素。实验结果与传统机器学习算法SVM、逻辑回归，集成学习算法AdaBoost进行对比。研究证明了XGBoost算法在慕课翘课预测中的有效性，不同类型课程影响因素的差异性。其结果对平台留存率、资源有效使用率提升具有重要现实意义。

关键词

慕课，翘课预测，XGBoost

1. 引言

共享经济是“互联网+”时代的新型服务模式和商业模式。近年来，共享经济在全球范围内产生了巨大的反响。共享更像是一种不可逆转的潮流趋势，从实体到服务、从线上到线下，渗透到人类生活的角角落落[1]。据普华永道报告显示，2025 年，全球共享经济产值或可达到 2300 亿英镑。

慕课(Massive Open Online Courses, MOOCs)，即大规模在线开放课程，被认为是共享服务经济领域“互联网+教育”的杰出代表。慕课或许能打破传统教育的桎梏，并将改变未来的教育领域[2]。数据表明，慕课正在促进高等教育转型并为科学研究提供有力支持[3]。

然而，近年来随着慕课的推广和深入，问题频频出现。其中，慕课完成率成为一大聚焦点。有相关研究表明，虽然完成率受很多因素影响[4]，但平均完成率非常低，在 6%左右[5]。考虑到平均完成率可能会受到个别课程影响而降低，但大多数课程的完成率也在 10%以下[6]。故探究学员行为影响因素，从而提升课程留存率、完成率有助于教育资源充分利用。预测学员是否有某种行为是典型的二分类问题。

XGBoost [7]是一种基于梯度提升决策树的集成学习算法。一经提出便在众多竞赛项目里大放异彩。XGBoost 因为出色的学习效果、快速的迭代速度以及可并行性等优势受到广泛关注并很快在各大领域得以应用。

现有的慕课预测研究主要以传统机器学习模型为主。卢晓航[8]等使用逻辑回归、SVM、LSTM 等对辍学进行预测，探究影响辍学的因子。王雪宇等[9]使用多元线性回归模型对学员辍课进行预测。Li W 等建立基于行为特征的新型多视图半监督学习模型，用于提高辍学预测准确率[10]。但以上只是基于慕课数据进行预测或提高预测准确度，并没有对慕课进行课程划分，探索不同种类课程影响因素的不同。此外，叶倩怡等[11]将 XGboost 用于商业销售预测，取得了较好效果。崔艳鹏等[12]提出一种基于 XGBoost 算法的 Webshell 检测方法，结果表明该算法优于单一的 Webshell 检测方法。这些均表明 XGBoost 在预测分析上有较好表现。

本文使用慕课理论、实操两类课程的后台数据。并将 XGBoost 应用于预测模型。通过挖掘学员行为、属性影响因素，建立二分类预测模型。从而探究学员翘课的影响因素，以及不同影响因素对于不同类型

课程的贡献程度。结果表明，与传统机器学习模型以及集成学习模型相比较，XGBoost 具有更高的预测准确度。且不同类型课程学员翘课影响因素存在较大差异。结果可用于提升平台留存率，提高慕课资源利用率。

2. 数据预处理

2.1. 数据来源

本课题研究的数据来源于数据来源于中国大学慕课网(https://www.icourse163.org/)《教师如何做研究》、《交互式电子白板教学应用》两门课程的后台数据，信息通过用户表、课程表、成绩表、学习日志表等16 个数据表来呈现。后台数据所包含的用户超过 1.5 万人，日志信息累计超过 100 万条。故数据表中包含的信息能很好的表现用户自身信息、用户交互行为。其中，《教师如何做研究》一课代表理论类课程，而《交互式电子白板教学应用》一课代表实操类课程。本文使用这些数据来建立模型预测学员翘课行为，从而探究不同类型课程翘课影响因子的差异性。

2.2. 特征选择

以往的研究中，学者多采用交互行为对学员流失进行预测。交互行为是客观可观察的，且数据是易于获取的。交互行为主要分为两个部分：课程交互和学员交互。课程交互主要指学员在学习过程中与课程资料、作业、测试等产生的交互，如卢晓航等[8]使用“视频点击流”、“课堂测验”作为指标对学生流失行为进行预测；贺超凯等[13]使用“学习事件次数”、“学习章节数”等进行学员学习行为分析。学员交互主要指学习过程中与其他学习者产生的交互行为，如论坛行为。Xing [14]、Fei M 等[15]均以“论坛发帖次数”为指标对学员流失进行探究，并验证了其显著性。部分研究以学员自身属性如“年龄”为指标，但其他基于自身属性的指标选取较少。

结合以往研究与本文数据集特点，选取了 3 类共计 19 个特征指标，对学员进行探究，来预测翘课情况进而研究不同类型课程的翘课影响因子的差异性。特征选取如表 1 特征描述：

Table 1. Description of factors
表 1. 特征描述

特征类型	特征	含义
课程交互	视频占比	学习中视频所占比例
课程交互	课程完成率	课程完成资源占总资源比例
课程交互	学习小节数	学习了多少个小节
课程交互	学习次数	一共学习多少次
课程交互	学习天数	一共学习多少天
课程交互	作业分	作业得分
课程交互	测试分	测试得分
课程交互	总学习时长	有效学习时长
学员交互	发帖数	发帖数量
学员交互	讨论数	讨论次数
学员交互	讨论分	依据讨论次数、质量所得分数
学员交互	评论数	评论他人帖子次数

Continued

学员交互	被回复数	评论被回复的次数
学员交互	被赞数	被赞次数
自身属性	年龄	学员年龄
自身属性	平台使用次数	学员初次注册到报名该课程的时长
自身属性	总发呆学习率	用户发呆学习比例
自身属性	是否发呆学习者	用户是否发呆学习者
自身属性	平均延迟学习天数	课程小节发布到学员学习平均时长

2.3. 数据预处理

数据预处理原理图如图 1 数据预处理流程图。用户通过多个终端访问服务器进行交互行为，从而产生数据池。我们从数据库中提取所需数据文件，进行数据清洗如多表连接、数据标注、特征选取等。从而产生原始数据文件。之后，又对原始数据文件进行缺失值填充、标准化、均衡化等处理从而得出能进行机器学习的“干净的”数据。

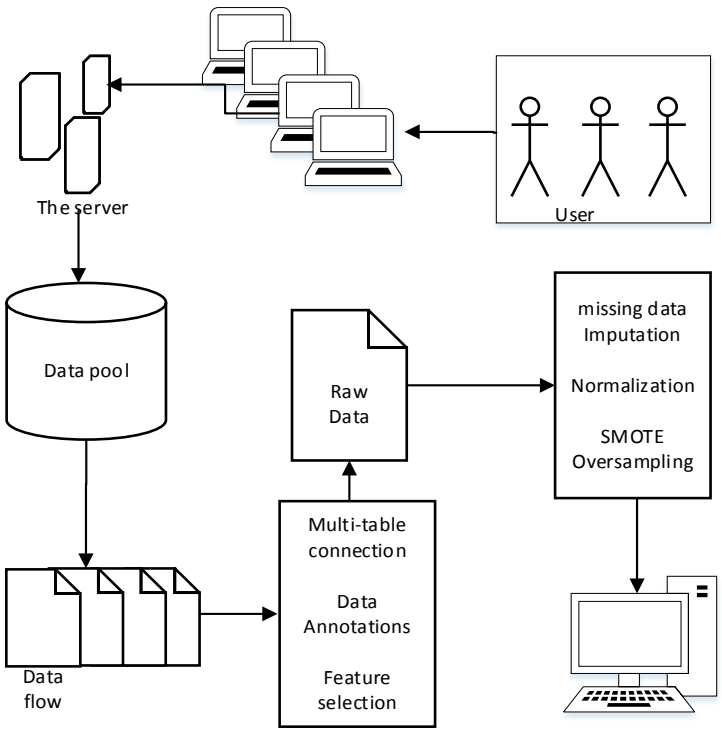

Figure 1. Data preprocessing schematic
图 1. 数据预处理流程图

1) 特征选取、计算及数据标注

原始数据并无直接的标签以及模型特征，需要了解每张表包含的数据意义，统一数据口径，将非结构化数据结构化并进行特征选取、计算，并对预测变量翘课与否进行标注。表 1 特征描述显示了特征选取后的结果。图 2 学员发呆行为判别流程图、图 3 学员翘课行为判别流程图对于分别不便理解的学员发呆行为和翘课行为进行定义。

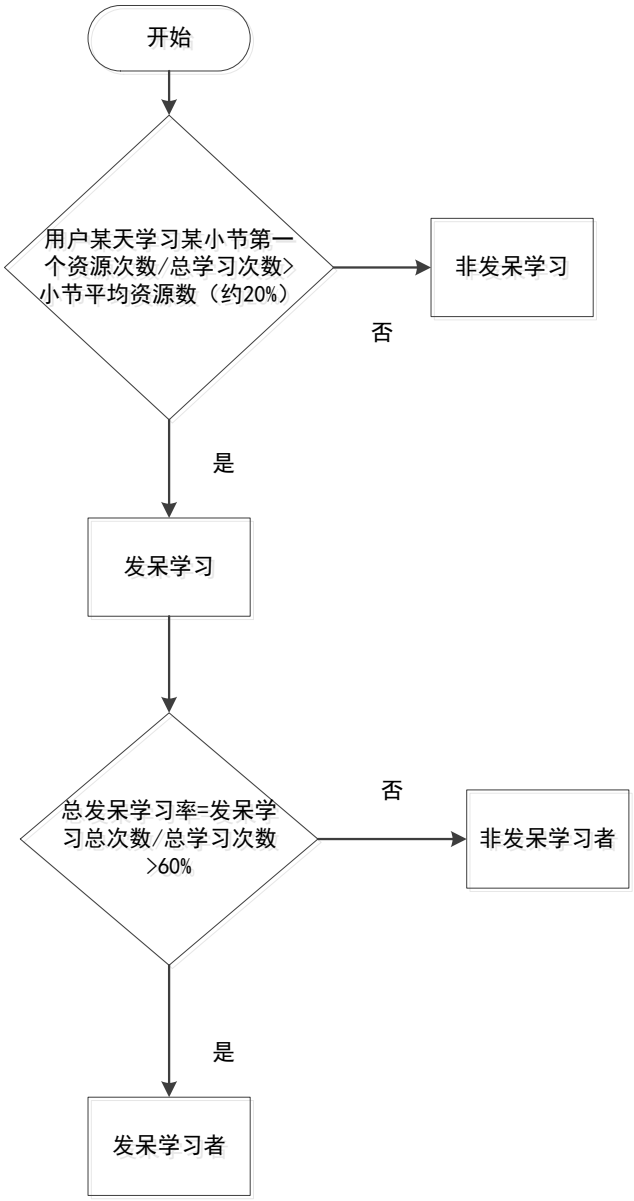

Figure 2. Student's dashed behavior discrimination flow chart
图 2. 学员发呆行为判别流程图

因为点击某小节时默认打开小节第一个资源，这时会存在用户打开但并没有进行学习的情况。我们想要探究这种情况对学员翘课的影响。

2) 缺失值填充

缺失值的存在会影响建模和预测质量，缺失值处理方式有删除、填充 0、填充均值、中位数等。采用何种方式需要进行衡量。本文根据缺失值填充对于研究的有无意义为根本，综合考虑简洁性、科学性等，对缺失值进行处理。

由于无学习记录的用户对本文无意义，本文对于缺失发呆用户属性的用户进行删除。对于年龄采用平均值填充。此外由于存在一些噪声和脏数据，对于重复注册用户予以删除。

3) 归一化处理

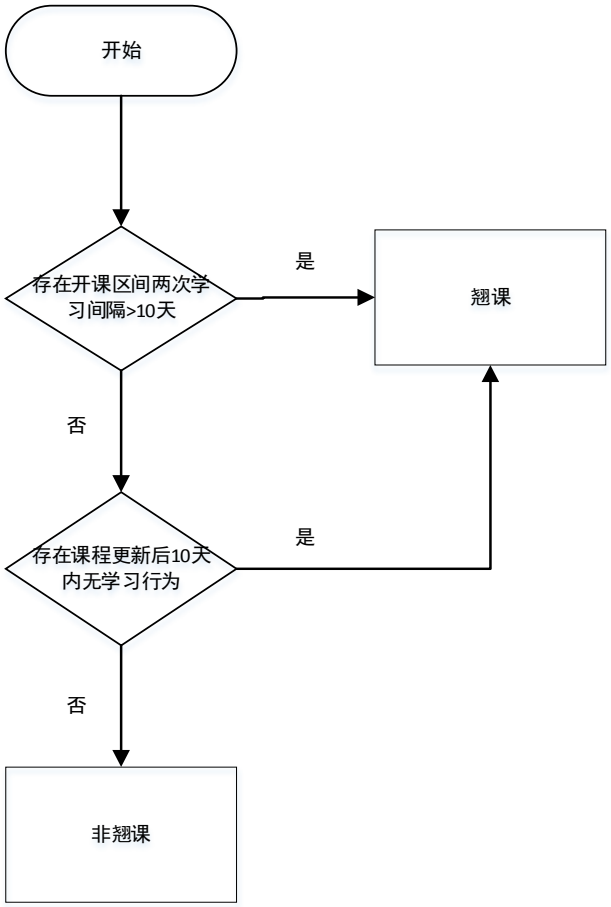

Figure 3. Student skipping behavior discriminating flow chart
图 3. 学员翘课行为判别流程图

由于数据集中某些属性值量纲不同，而消除量纲有利于加速优化过程。故拟对数据进行归一化。归一化一般有两种：① 最值归一化：将所有数据映射到 01 之间；② 均值方差归一化：数据处理为均值 0，方差 1 的分布中。但考虑到后期进行多次训练测试数据分离，对数据进行统一的均值方差归一化会使训练集受测试集影响，故采用归一化处理。同时为了纠正误差，对归一化之后的数据+0.00001.

算法 1. 归一化伪代码

Normalization Algorithm

Begin

1: Obtain factor

2: Get range R = factor.max()-factor.min()

3: NormalizaFactor_i = (factor-factor.min())/R

4: Correct the errors New_factor = Factor_i+0.000001

End

4) 非平衡化处理

大多数机器学习算法在数学原理上都假定数据均匀分布。数据集不均衡会对机器学习算法的分类效

果带来极大的负面影响。本文的目标在于解决慕课中高翘课率的问题，数据集是不均衡的。《交互式电子白板教学应用》数据集中翘课人数：不翘课人数约为 7.2，《教师如何做研究》约为 3.2。这会对建模结果产生很大影响。

解决非平衡问题有两种方法即减少多数类别样本的欠采样和增加少数样本的过采样。本文采用过采样经典算法 SMOTE，对训练数据集进行处理。

算法 2. SMOTE 过采样伪代码

SMOTE Oversampling Algorithm

Begin

```
1: defover_sampling(T, N, k): // T 为少数类样本个数，N%为采样比例，k 为 k 近邻//
N = int(N/100)
Fori = 1 to T
Nnarray = neighbors.kneighbors(i,k) //对于每个少数类样本，求其 k 近邻//
Populate (N, i, nnarray)
end for
2: defpopulate (N, i, nnarray):
Nn = random 1-k //根据采样比例在 k 近邻中随机选择//
while N! = 0
for a = 1 to num //对 nn 中的每个近邻，求新样本//
dif = samples[nnarray[nn]]-samples[i] //求新样本方向以及 k 近邻与原样本距离//
gap = random 0-1
synthetic[newindex] = samples[i]+gap*dif //得出新样本//
newindex+ = 1; N--
end for
end while
```

End

3. 基于 XGBoost 的翘课预测

3.1. XGBoost 算法

XGBoost (Extreme Gradient Boosting)即极度梯度提升树[16]，相比于一般的梯度提升树(gradient boosted tree, GBDT)，其优势在于避免了过度拟合、泛化性能优；加快优化速度，减小内存消耗；在特征上并行处理。

目标函数由训练误差和正则化项构成。训练误差是为了衡量模型的预测能力，而正则化项是为了控制模型复杂度，避免过拟合。其中 $l\left(y_i,\hat{y}_t\right)$ 表示样本 i 的训练误差，$\Omega\left(f_k\right)$ 表示树的复杂度。目标函数：

$$J(\theta)=\sum_{i=1}^{n} l\left(y_i,\hat{y}_t\right)+\sum_{k=1}^{K}\Omega\left(f_k\right) \tag{1}$$

XGBoost 为增量训练，即在前一步的基础之上新增一棵树，并优化新增树。

$$\begin{aligned}
&\hat{y}_i^{(0)}=0\\
&\hat{y}_i^{(1)}=f_1\left(x_i\right)=\hat{y}_i^{(0)}+f_1\left(x_i\right)\\
&\hat{y}_i^{(2)}=f_1\left(x_i\right)+f_2\left(x_i\right)=\hat{y}_i^{(1)}+f_2\left(x_i\right)\\
&\quad\vdots\\
&\hat{y}_i^{(t)}=\sum_{k=1}^{t} f_k\left(x_i\right)=\hat{y}_i^{(t\text{-}1)}+f_t\left(x_i\right)
\end{aligned} \tag{2}$$

这里，第 t 步中新增的树，是在前一步基础上使目标函数最优的树：

$$\hat{J}_i^{(t)} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)} + f_t\left(x_i\right)\right) + \Omega\left(f_t\right) \tag{3}$$

使用泰勒展开公式来逼近，其中，$g_i = \partial_{\hat{y}_i^{(t-1)}}\left(y_i, \hat{y}_i^{(t-1)}\right)$，$h_i = \partial_{\hat{y}_i^{(t-1)}}\left(y_i, \hat{y}_i^{(t-1)}\right)$

$$J^{(t)} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)} + g_i f_t\left(x_i\right) + \frac{1}{2} h_i f_t^2\left(x_i\right)\right) + \Omega\left(f_t\right) \tag{4}$$

不考虑常数项，则目标函数为

$$J^{(t)} = \sum_{i=1}^{n} l\left(g_i f_t\left(x_i\right)\right) + \frac{1}{2} h_i f_t^2\left(x_i\right) + \Omega\left(f_t\right) \tag{5}$$

在 XGBoost 中，T 表示叶子节点数，w 表示叶子节点数值，则复杂度定义为：

$$\Omega\left(f\right) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \tag{6}$$

则

$$J^{(t)} = \sum_{j=1}^{T}\left[G_j w_j + \frac{1}{2}\left(H_j + \lambda\right) w_j^2\right] + \gamma T \tag{7}$$

其中，$G_j = \sum_{i \in I_j} g_i$，$H_j = \sum_{i \in I_j} h_i$

对 w_j 求导可得，

$$w_j^* = -\frac{G_j}{H_j + \lambda} \tag{8}$$

$$J^* = -\frac{1}{2}\sum_{j=1}^{T} \frac{G_j^2}{H_j + \lambda} + \gamma T \text{ 。} \tag{9}$$

3.2. 评价标准

本文使用多个指标对模型进行评价，包括 recall、precision、F1、ACC、AUC 等。首先引入混淆矩阵如表 2 混淆矩阵：

Table 2. Confusion_matrix
表 2. 混淆矩阵

数量	预测为正样本	预测为负样本
标签为正样本	TP (True Positive 对的正样本)	FN (False Negative 错的负样本)
标签为负样本	FP (False Positive 错的正样本)	TN (True Negative 对的负样本)

精确率(precison)即查准率，即计算预测出的正样本中，有多少是正确分类：

$$P = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{10}$$

召回率(recall)即查全率，即真正的正样本中，有多少被正确预测：

$$R = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{11}$$

F1：精确率与召回率的调和平均值：

$$\mathrm{F1}=\frac{2*P*R}{P+R} \tag{12}$$

精度(ACC)：被正确分类的样本占总样本的比:

$$A=\frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}} \tag{13}$$

AUC (Area under curve)：M 为正样本个数，N 为负样本个数时

$$\mathrm{AUC}=\frac{\sum_{\text{正样本}} rank(score)-\frac{M*(M+1)}{2}}{M*N} \tag{14}$$

综上，使用这些评价标准可以从多个角度更准确地衡量分类结果。

3.3. 实验流程

实验主要流程将上一小节处理好的数据进行训练、测试集划分。进入不同模型进行训练、预测并进行结果对比，并对 XGBoost 模型进行参数优化，从而对最终的模型进行测试并对结果加以分析。具体实验流程如图 4。

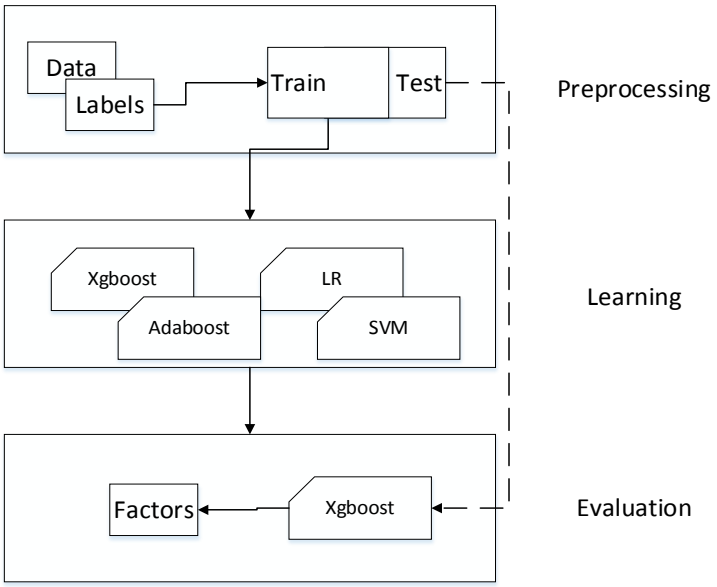

Figure 4. Experimental flow diagram
图 4. 实验流程原理图

4. 实验与结果分析

4.1. 实验

实验使用内存为 8GB DDR4，2400MHz，i7-7500U 处理器笔记本一台。使用 Win10 操作系统。Python 为 3.6 版本。

本文最后模型所用的数据包括《教师如何做研究》6430 条记录、《交互式电子白板教学应用》9746 条记录，包括 18 个特征和一个预测值即翘课与否。这些数据涵盖两门课程多个学期超 15000 名学员的日志信息。本文将数据以 7:3 进行训练测试划分并调用 python3 的 siki-learn 包进行了 XGBoost、AdaBoost、

SVM、逻辑回归的建模和预测，结果如表 3《交互式电子白板教学应用》不同模型预测效果对比、表 4《教师如何做研究》不同模型效果对比。

Table 3. Comparison of Different Models in "Interactive Whiteboard Teaching Application"
表 3. 《交互式电子白板教学应用》不同模型预测效果对比

	XGBoost	AdaBoost	SVM	LR
Recall	0.9069	0.8648	0.7962	0.7806
Precision	0.9336	0.8990	0.9207	0.8949
F1	0.9201	0.8816	0.8539	0.8338
Auc	0.9208	0.8832	0.8633	0.8438
Acc	0.9207	0.8830	0.8629	0.8434
时间	4.679 s	1.750 s	2.901 s	1.886 s

Table 4. Comparison of different models of "How do teachers do research"
表 4. 《教师如何做研究》不同模型效果对比

	XGBoost	AdaBoost	SVM	LR
Recall	0.7920	0.6499	0.6212	0.5438
Precision	0.7960	0.7475	0.6616	0.5936
F1	0.7940	0.6953	0.6408	0.5676
Auc	0.7974	0.7183	0.6563	0.5910
Acc	0.7975	0.7193	0.6568	0.5917
时间	2.574 s	1.817 s	4.113 s	1.799 s

从结果来看，XGBoost 在 recall、precision、AUC 上均有较好的表现。虽然在时间上稍有劣势，但从量级上看并没有很大的缺陷，可用于预测学员翘课。

为了预测的准确性，本文选取预测较差的《教师如何做研究》的模型对其进行参数优化。本文采取了网格搜索的方式，针对 F1 值，对 n_estimators、max_depth、min_child_weight、subsample、colsample_bytree、reg_alpha、gamma 进行了调优。并将模型用于《交互式电子白板教学应用》结果如表 5。

Table 5. The effect after adjusting the parameters of XGboost
表 5. XGBoost 调参后效果

	教师如何做研究	交互式电子白板教学应用
Recall	0.8623	0.9443
Precision	0.8233	0.9430
F1	0.8423	0.9437
Auc	0.8412	0.9432
Acc	0.8409	0.9432
时间	9.183 s	11.902 s

可以发现，调参后效果有显著提升，这为之后的对比分析提供了可信度支持。

通过 XGBoost 可以计算每个特诊对于模型的贡献程度从而确定哪些变量对于学员翘课行为的影响更

为显著，进而对不同类型课程进行对比分析。

4.2. 结果分析

首先，由图 5、图 6 可以看出对于慕课学习来说，课程交互、学员交互、自身属性均会对学员翘课产生一定的影响。和我们平常认知相同，学习越努力，课程完成度高，参与积极，互相协作的学员更不容易翘课。那么对于平台来说，想要使学员留存率提升，要及时触达课程完成率低、学习天数较少、与人交互较少的学员。可以通过邮件召回、连续签到有奖、互动加分等形式来完成。

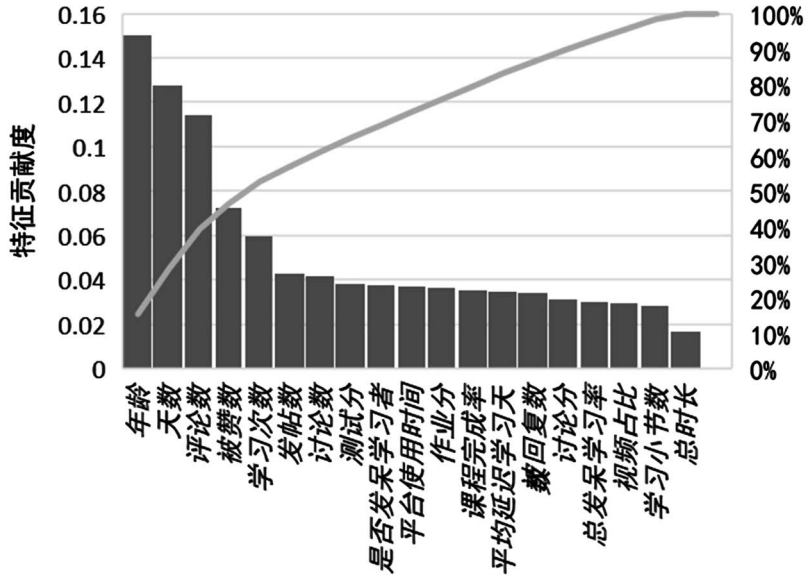

Figure 5. Feature contribution of "How do teachers do research"
图 5. 《教师如何做研究》特征贡献度

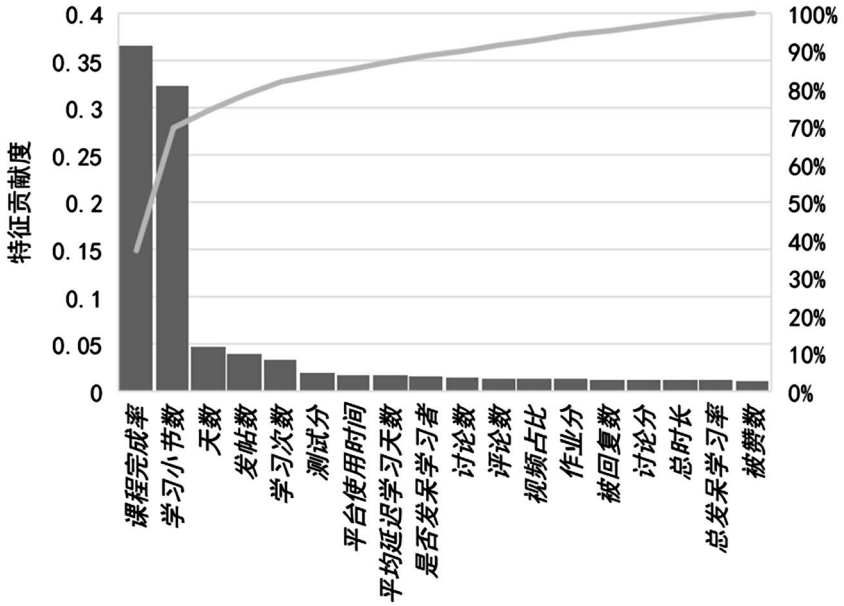

Figure 6. Feature Contribution of "Interactive Whiteboard Teaching Application"
图 6. 《交互式电子白板教学应用》特征贡献度

其次，研究表明，理论和实操类课程的翘课影响因子存在差异。对于《教师如何做研究》所代表的理论类课程，分别归属于自身属性、课程交互、学员交互的年龄、学习天数、评论数贡献度最大。且其他特征的贡献度较为均衡。这表明对于理论类课程，一个方面的影响不是翘课决定性因素，这也恰恰说明学好理论类课程需要对各个方面加以努力。至于年龄贡献度最大，也许是因为课程受众学员年龄较为密集。对于《交互式电子白板教学应用》所代表的实操类课程，课程完成程度即课程完成率和小节完成度占据了 70%以上的贡献度。这一方面可以说明实操类课程更需要和课程间的交互，才能使学员学有所得，产生学习的信心；另一方面其他特征贡献率较低可能源自实操类课程的特点，学习重要性，学员之间的交互可能加剧因课程交互低带来的不自信，从而产生翘课行为。对于这类课程平台可以多敦促学员的完成情况，或将完成情况加入课程考核中。

5. 结语

慕课作为共享经济+教育的成果，对于我国共享经济和教育发展有着重要作用。但慕课的高辍学率严重影响慕课资源的使用，从而对我国教育产业化、信息化发展，国际教育竞争力提升产生了严重阻挠。

目前学者对于慕课完成率、翘课率有一定的研究。但并未对理论性、实操类课程进行区分。没有根据两类课程的不同探究不同因素对慕课教学效果的影响不同，并得出提高课程效果的不同策略。

本文使用慕课后台数据结合集成学习算法中的 XGBoost 算法对学员翘课行为进行了建模、预测，打破了传统机器学习方法预测准确率低的问题。并将一些新的指标应用于预测模型，提升了模型的可解释性。此外，本文对不同类型课程的研究因素进行了重要性分析和差异性分析。确定了不同课程的影响因子贡献程度。该研究有助于学员自身学习提升以及平台留存率提升。对于慕课这一共享服务经济有着重要的现实意义。

参考文献

[1] 富切尔·博茨曼, 路·罗杰斯. 共享经济时代: 互联网思维下的协同消费商业模式[M]. 唐朝文, 译. 上海: 上海交通大学出版社, 2015.

[2] Stein, L.A. (2012) Casting a Wider Net. *Science*, **338**, 1422-1423. https://doi.org/10.1126/science.1230710

[3] Waldrop, M.M. (2013) Online Learning: Campus 2.0. *Nature*, **495**, 160-163. https://doi.org/10.1038/495160a

[4] Yousef, A.M.F., Chatti, M.A., Schroeder, U., *et al*. (2014) What Drives a Successful MOOC? An Empirical Examination of Criteria to Assure Design Quality of MOOCs. 14*th International Conference on Advanced Learning Technologies*, Athens, 7-10 July 2014, 44-48. https://doi.org/10.1109/ICALT.2014.23

[5] Jordan, K. (2014) Initial Trends in Enrolment and Completion of Massive Open Online Courses. *International Review of Research in Open and Distance Learning*, **15**, 133-160. https://doi.org/10.19173/irrodl.v15i1.1651

[6] Hone, K.S. and Said, G.R.E. (2016) Exploring the Factors Affecting MOOC Retention: A Survey Study. *Computers & Education*, **98**, 157-168. https://doi.org/10.1016/j.compedu.2016.03.016

[7] Chen, T.Q. and Guestrin, C. (2016) XGBoost: A Scalable Tree Boosting System. *Proceedings of the* 22*nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, 13-17 August 2016, 785-794. https://doi.org/10.1145/2939672.2939785

[8] 卢晓航, 王胜清, 黄俊杰, 陈文广, 闫增旺. 一种基于滑动窗口模型的 MOOCs 辍学率预测方法[J]. 数据分析与知识发现, 2017, 1(4): 67-75.

[9] 王雪宇, 邹刚, 李骁. 基于 MOOC 数据的学习者辍课预测研究[J]. 现代教育技术, 2017, 27(6): 94-100.

[10] Li, W., Gao, M., Li, H., *et al*. (2016) Dropout Prediction in MOOCs Using Behavior Features and Multi-View Semi-Supervised Learning. *International Joint Conference on Neural Networks*, Vancouver, 24-29 July 2016, 3130-3137. https://doi.org/10.1109/IJCNN.2016.7727598

[11] 叶倩怡, 饶泓, 姬名书. 基于 XGBoost 的商业销售预测[J]. 南昌大学学报(理科版), 2017, 41(3): 275-281.

[12] 崔艳鹏, 史科杏, 胡建伟. 基于 XGBoost 算法的 Webshell 检测方法研究[J]. 计算机科学, 2018, 45(S1): 375-379.

[13] 贺超凯, 吴蒙. edX 平台教育大数据的学习行为分析与预测[J]. 中国远程教育, 2016(6): 54-59.

[14] Xing, W., Chen, X., Stein, J., *et al*. (2016) Temporal Predication of Dropouts in MOOCs: Reaching the Low Hanging Fruit through Stacking Generalization. *Computers in Human Behavior*, **58**, 119-129. https://doi.org/10.1016/j.chb.2015.12.007

[15] Fei, M. and Yeung, D.Y. (2015) Temporal Models for Predicting Student Dropout in Massive Open Online Courses. *IEEE International Conference on Data Mining Workshop*, Atlantic City, 14-17 November 2015, 256-263. https://doi.org/10.1109/ICDMW.2015.174

[16] Chen, T. and Tong, H. (2014) Higgs Boson Discovery with Boosted Trees. *International Conference on High-Energy Physics & Machine Learning*, Montreal, 13 December 2014, 69-80.