

面向 Canvas 学习管理系统的 在线数据分析和挖掘研究^{*}

闫宏飞^{1,2} 胡扬¹ 王彦恺¹

(1. 北京大学计算机学院, 北京 100871;

2. 北京大学大数据科学研究中心, 大数据分析与应用技术国家工程实验室, 北京 100871)

摘要:[目的/意义] 新冠疫情带来线上教学以及线上线下混合教学的普及, 然而教师和学生仍然普遍认为线上教学的效果弱于线下教学。因此, 本文旨在分析两种教学方式的根本区别, 以及探讨如何充分发挥线上教学的潜能, 使学生拥有最好的学习体验。[方法/过程] 本文选取“计算概论(B)”本科课程的线上线下两个平行班, 深度挖掘教学管理系统可获取的学生行为数据, 提取对学生学习成绩的关键影响因素, 采用机器学习方法做出了重要性分析, 并对学生课程的最终成绩做出预测。[结果/结论] 本文认为, 影响学生学习效果的关键性因素包括学生的平时表现、投入程度、互动情况、社交关系和自我评价。两种教学方式的区别主要体现在课程投入程度和互动情况方面。最后结合实际, 从学习效果评估和优化线上教学设计两个方面提出教学建议。

关键词: 回归分析 教学管理系统 在线教学

分类号: G251

DOI: 10.31193/SSAP.J.ISSN.2096-6695.2022.01.06

0 引言

新冠肺炎疫情发生以来, 全国中小学、高校等都推迟了开学时间。因此, 教育部发出利用网络平台, 展开“停课不停学”的倡议^[1]。一时间, 在线课程、网络直播、视频会议等在线教育实践, 在全国各类学校得到广泛展开。通过2020年春季一学期的课程实践与探索, 可以看到, 在线教学和线上线下混合教学作为高等教育的一种新形态不仅可行, 而且已经逐渐成为教育模式

^{*} 本文系北京大学“教育大数据研究项目”(项目编号: 2020YBC01)的研究成果之一。

[作者简介] 闫宏飞 (ORCID: 0000-0001-5914-8585), 男, 副教授, 博士, 研究方向为信息检索、计算金融, Email: yanhf@pku.edu.cn; 胡扬 (ORCID: 0000-0003-4730-4725), 女, 硕士研究生, 研究方向为知识图谱, Email: 1901213321@pku.edu.cn; 王彦恺 (ORCID: 0000-0002-3258-1757), 男, 硕士研究生, 研究方向为机器学习, Email: 2001213038@pku.edu.cn。

的新常态^[2]。

在此之前, 线上教育主要作为线下教学的一种辅助形式, 或以大规模开放线上课程 (MOOC) 的形式体现^[3]。诚然, 线上教学有其广泛的应用前景, 但以往的 MOOC 类课程面向全部互联网参与者授课, 学生参与感和互动感较差, 教师对学生也几乎起不到课外的监督指导作用, 与传统线下教学相比, 其授课效果局限性非常明显。不过, 线上授课也有其得天独厚的优势, 除了学生的参与时间更灵活自由之外, 线上教学平台能够大批量收集学生的学习数据, 并利用这些数据对学生的行为做出分析甚至预测, 而在线下课程中, 这些信息的收集完全依赖教师和班长进行人工采集^[4]。

目前, 授课者面临的挑战是如何利用线上教学的形式, 取其长补其短, 即在保持线上教学灵活性的同时, 优化课堂互动和授课模式, 加强学生参与感和投入度, 并有效利用教学平台采集的信息, 对学生的学习过程进行监督和指导, 最终让学生收获更好的学习成果。

针对该挑战, 本文选取北京大学 2020 年秋季学期本科生“计算概论 (B)”课程, 通过教学管理系统以及授课过程中积累的学生参与行为、平时表现、自我评价、群聊互动、社交关系等方面的特征, 探索不同特征对学生的期末成绩的影响, 并做出重要性分析, 再根据实验结果并结合本课程教学过程, 给出未来教学建议^[5,6]。

此外, 目前虽然有共识认为线上教学效果较差, 但是仍然没有较为深入定量的分析, 来解释为什么线上教学效果差, 以及如何利用这些解释来优化线上教学体验^[7]。针对这一问题, 由于“计算概论 (B)”课程分为两个班级, 两班的教学内容和进度相同, 但是分别采用线上和线下方式授课, 线下班级面向在校学生, 线上班级面向因疫情无法返校的学生^[8]。本文通过两个班级的对比实验, 对线上、线下两种教学模式的特性和优劣做了更为深入的分析, 可以作为后续教学设计及创新的参考^[9]。

最后, 本文使用机器学习模型对学生成绩做出预测, 此模型同时适用于线上和线下班级的学生。期望能在以后的教学过程中, 将此模型用来实时评估学生的学习成果, 对学习效果差的学生进行单独辅导, 或者根据学生学习效果适当调整课程难度等^[10,11]。

1 数据获取

北京大学 2020 年秋季学期“计算概论 (B)”课程, 是面向本科一年级非计算机专业学生的编程基础课, 分为两个班级: 12 班是线下班 (共 127 人), 13 班是线上班 (共 37 人)。两班教学内容和进度相同, 线下班级面向在校学生, 线上班级面向因疫情无法返校的学生。两个班级课程都在 Canvas 教学管理平台上展开, Canvas 上的内容包括 16 个单元课件、17 次作业、16 次每次 3 小时的课程回放视频。除 Canvas 外, 本课程还建了两个微信群, 用作日常交流和答疑, 随着课程进展进行了 9 次问卷调查, 并收集了学生在班级内的社交关系图。课程主要练习平台是 codeforces.com (以下简称“CF”) 和 cs101.openjudge.cn (以下简称“OJ”)。

本文所用数据来源主要分为以下几类: Canvas 教学管理平台、课程调查问卷、平时表现加分及期末机考通过题数、课程微信群聊天记录、学生班内社交关系情况。收集到的所有信息列表如表 1 所示。

(1) Canvas 教学管理平台。在线教学依赖于教学管理系统 (Learning Management System, 简称 LMS), 一般来说, 教师和学生都有 LMS 账号, 教师创建课程, 学生加入课程^[12]。教师和学生系统中的所有行为都会被记录, 课程相关的一切数据也都需要从系统获取。本研究关注的 2020 年秋季“计算概论 (B)”课程在北京大学的 Canvas 教学管理平台上开展, 因此可以从 Canvas 平台上获取两班每位同学的全部 17 次作业分数、作业总分、作业按时提交比例、上次参与时间、上次登出时间、全部页面浏览、全部参与次数, 这些特征全部为平台自动统计。

(2) 课程调查问卷。随着课程推进, 为了能实时了解学生的学习状态和反馈, 总共对两个班学生进行了 9 次问卷调查, 囊括了选课、期中、期末这些关键时间节点。问卷调查内容包括: 基本信息 (12 项)、计算机原理和编程实践的具体学习情况 (10 项)、算法相关内容 (6 项)、总体学习情况 (4 项)。

(3) 平时表现加分及期末机考通过题数。这部分信息由助教收集。平时表现加分最多为 3 分, 主要考虑以下几种情况: 作业被当成优秀范例展示; 为其他同学答疑; 参与线下答疑讨论以及其他情况 (例如做了一个额外的 project)。期末机考通过题数由 OJ 考试平台导出, 由于期末成绩主要由机考成绩决定, 因此, 期末机考通过题数可以看成学生最终的期末成绩, 或学习效果的体现。

(4) 课程微信群聊天记录。两班有各自的微信群, 同学们讨论交流也比较活跃 (聊天记录 12 班 555MB, 13 班 384MB)。通过微信的 message、rcontact、chatroom 这些数据表, 以 CSV 格式导出两个微信群的所有聊天记录, 其中, 12 班学生的微信讨论较多, 共有 13216 条, 而 13 班只有 3248 条。针对这部分信息, 对每个学生的微信发言次数 wechat_freq 进行统计, 并作为其微信群活跃度的体现。

(5) 学生班内社交关系情况。通过问卷调查的形式收集每位学生在班内熟悉的同学, 并据此构建了学生社交关系图, 如图 1 所示。选取学生的朋友数目 friends_cnt 和朋友的期末平均通过数 friends_score (也就是每个节点的度和节点的邻居节点的平均分数) 来代表学生社交情况。

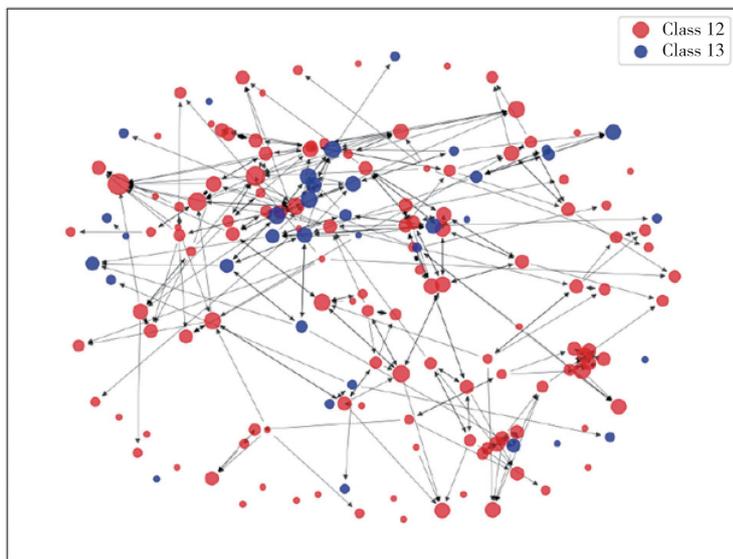


图 1 学生社交关系图

2 数据处理

由于获取到的特征来源复杂, 且特征数目较多, 需要对数据做清洗和预处理, 才能用于之后的回归模型^[13]。数据处理主要包括以下几个方面, 最终使用的特征如表 1 所示。

(1) 特征选择: 首先, 由于作业次数较多, 因此选择用所有作业的总分来代表作业成绩; 其次, 由于除了化学学院外, 其他学生所属学院较为分散, 所以暂时去掉院系特征。

(2) 分类特征哑变量处理: 一般来说, 分类变量需要处理成哑变量。分类变量可以分为名义的 (nominal) 和序数的 (ordinal) 两种, 其中, 名义的表示类别之间没有大小关系, 如性别、学生所属班级, 都是没有相对大小关系的; 序数的表示类别之间有大小关系, 比如对某个说法打分, 第一个选项是低于 5 分, 第二个选项是高于 5 分。这里将性别和班号这类名义型分类变量处理成哑变量, 而将序数型分类变量作为连续变量处理。

(3) 数值特征归一化处理: 对连续变量, 一般需要做归一化处理, 以保证所有特征的范围相差不要太大, 这里选择 MinMaxScaler, 把数据处理到 0 ~ 1 的范围内。MinMaxScaler 的计算过程用公式表示为:

$$X_{scaled} = X_{std} * (\max - \min) + \min$$

(4) 缺失值填充: 一般来说, 数据中会存在缺失值, 这些缺失值会影响后续的回归模型的计算, 因此需要对缺失值进行填充, 在此选用“0”来做填充。

(5) 去除特征共线性: 在多元线性回归模型中, 一般要保证特征之间没有共线性, 否则会导致数值解不稳定, 而且难以解释模型结果。一般使用 variance inflation factor (VIF) 来衡量自变量特征之间的共线性程度, VIF 值越大, 说明该特征和自变量中的其他特征有共线性关系, 需要移除。所以作为数据预处理的最后一步, 使用 VIF 来衡量特征之间的共线性程度, 并且递归地去除 VIF 值大于 10 的特征, 以去除有共线性关系的特征。

表 1 所有特征列表

特征来源	全部特征	选入模型的特征
学生基本信息	姓名 学号 年级 性别 系所专业 教学班号 中期退课状态	年级 (class_level) 教学班号 (class_num)
自我评估问卷调查	手机号 是否教室听课 是否直播听课 是否回看课堂视频听课 编程竞赛基础 计算机原理	直播听课 回看课堂视频听课 教室听课 编程竞赛基础 (oi_level) 每周投入时间 (weekly_hours) OJ 通过的练习题数目 (oj_ACs)

续表

特征来源	全部特征	选入模型的特征
自我评估问卷调查	每周投入时间 OJ通过的练习题目数 是否觉得自己的计算机思维有所提升 是否觉得自己的数学思维有所提升 是否感觉学到了很多新知识 是否觉得时间的投入是值得的 模拟考通过题目数	计算机原理 (principle of computer) 是否感到学到了新知识 (learned_knowledge) 是否感到投入是值得的 (effort_worthwhile) 是否数学思维有提升 (mathematical_thinking) 是否计算机思维有提升 (cs_thinking) 模拟考试通过题数 (mock_exam_result)
Canvas系统内学生行为	作业按时提交比例 上次参与时间 上次登出时间 全部页面浏览 (Page View) 全部互动次数 (Participation)	作业按时提交比例 (Assignment on time percent) 全部页面浏览 (Page Views) 全部互动次数 (Participations)
Canvas系统内平时成绩	17次作业每次的分数 全部作业总分	全部作业总分 (Final Points)
微信讨论活跃度	微信发言次数	微信发言次数 (wechat_freq)
学生社交关系情况	朋友总数 朋友均分	朋友总数 (friends_cnt) 朋友均分 (friends_score)
平时表现加分	Bonus	Bonus
期末机考通过题数	ACs@FE	ACs@FE

3 学生成绩影响因素分析

在前一个章节的数据收集和处理部分,已经完成了特征的提取和预处理,在这一部分,主要利用这些特征,对两个班级全部学生整体进行分析,说明哪些特征对学生成绩影响较为显著。由于两个班级的学生量较少,因此选用OLS多元线性回归模型,通过特征系数来代表不同特征的重要性程度^[14]。使用学生期末通过题目数为因变量,其他特征为自变量,得到OLS结果见表2。

表2 对全部学生的OLS结果

	coef	std err	t	P> t	[0.025	0.975]
const	1.0852	1.004	1.081	0.282	-0.902	3.072
bonus	-0.4567	0.333	-1.371	0.173	-1.116	0.203
Final Points	1.9359	0.619	3.128	0.002	0.711	3.161
Assignment on time percent	0.8376	0.487	1.719	0.088	-0.127	1.802

续表

	coef	std err	t	P> t	[0.025	0.975]
Page Views	-1.2714	0.762	-1.669	0.098	-2.779	0.236
Participations	-2.0436	0.963	-2.122	0.036	-3.950	-0.138
Live class	0.1420	0.239	0.594	0.553	-0.331	0.615
Review live class	0.0482	0.239	0.201	0.841	-0.426	0.522
class_level	-0.7131	0.467	-1.528	0.129	-1.637	0.210
oi_level	-0.9732	0.426	-2.287	0.024	-1.816	-0.131
weekly_hours	0.5990	0.636	0.941	0.348	-0.661	1.859
oj_ACs	-0.4126	0.326	-1.266	0.208	-1.058	0.232
principle_of_computer	-0.4878	0.454	-1.074	0.285	-1.387	0.411
learned_knowledge	1.5631	0.722	2.165	0.032	0.134	2.992
effort_worthwhile	0.6585	0.582	1.132	0.260	-0.493	1.810
mathematical_thinking	-0.5414	0.457	-1.185	0.238	-1.445	0.363
cs_thinking	0.7244	0.467	1.550	0.124	-0.201	1.649
mock_exam_result	1.9890	0.340	5.848	0.000	1.316	2.662
wechat_freq	0.9847	0.574	1.716	0.089	-0.151	2.120
friends_cnt	-1.1751	0.596	-1.973	0.051	-2.354	0.004
friends_score	0.7553	0.468	1.613	0.109	-0.171	1.682
性别_男	0.1553	0.237	0.656	0.513	-0.313	0.624
class_num_1	-0.1335	0.506	-0.264	0.792	-1.134	0.867

在这些系数里面，只需要关注经过 t 检验显著的那些，也就是 $P > t$ 这一列为 0.1 以下或者略微超过 0.1 的部分。可以观察到，作业总分 (Final Points)、作业按时提交比例 (Assignment on time percent)、Canvas 页面浏览 (Page Views)、Canvas 互动情况 (Participations)、竞赛基础 (oi_level)、认为自己学到了知识 (learned_knowledge)、模拟考分数 (mock_exam_results)、微信讨论活跃度 (wechat_freq)、朋友数目 (friends_cnt)、朋友的平均分 (friends_score)，这些特征对学生成绩的影响是显著的。对所有影响显著的特征，画出其特征重要性图，如图 2 所示。

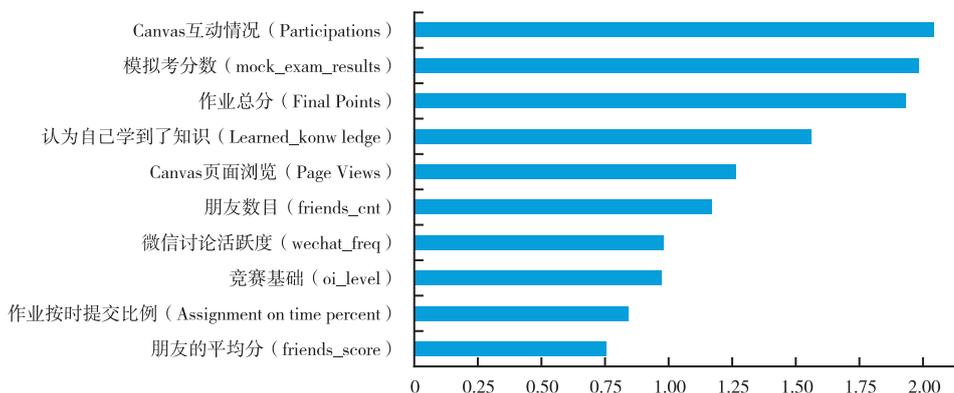


图2 特征重要性图

其中, Canvas 页面浏览 (Page Views)、Canvas 互动情况 (Participations) 和朋友数目 (friends_cnt) 的影响是负向的 (oi_level 虽然系数是负的, 但是实际含义还是竞赛基础越好, 成绩就越高), 初步观察会得出课程参与相关的指标对成绩的影响是正向的结论。但是对本门课程来说, Canvas 页面浏览 (Page Views)、Canvas 互动情况 (Participations) 有负向影响是因为 Canvas 并不是唯一的互动平台、甚至是比较次要的互动平台, 学生可以更多地通过线下答疑, 或者在微信群讨论等方式参与课程互动。因此, 在 Canvas 上的浏览和互动不能代表学生的课程互动参与整体情况, Canvas 上互动越多的人, 不一定总的互动行为越多。

为了验证 Canvas 上互动不代表学生总的互动情况这一猜想, 画出 Canvas 页面浏览 (Page Views)、Canvas 互动情况 (Participations) 和学生微信互动次数 (wechat_freq) 的关系图, 如图 3 所示。为了表示清楚, 对三种特征都取了 LOG 处理, 且图 3 中点越大, 代表学生期末通过题目数越多。从图 3 可以看出, 学生在 Canvas 上的互动行为分布较为集中, 这是因为 Canvas 用于交作业, 大部分人提交作业的次数都是一样的, 在 Canvas 上讨论区的发言或评论也很少。图 3 越往右的部分, 面积较大的点越密集, 也就是成绩高的学生越密集, 而最左边的部分几乎都成绩较差, 可以初步反映微信活跃度对成绩的影响。而从横向取一个切片, 观察发现互动较大或较小, 点的横向分布均没有什么特点, 也就是说, 看起来互动较多的人和较少的人, 并没有在微信发言频次上有明显的差异, 同样的结论对页面浏览也成立, 说明这两点 Canvas 上的互动统计, 并不能代表学生的真实课程互动情况, 如果需要利用 Canvas 收集的数据, 或许需要更细粒度的特征提取。

而朋友数目 (friends_cnt) 的负向影响是不符预期的。通过回看问卷问题发现, 此处要求学生填写 3 个熟悉的同学姓名 (只有 3 个位置可以填), 只有少部分人填写的熟悉同学少于 3 个, 然而有一大部分同学填写了超过 3 个 (有的甚至写了 6 个)。从数据中可以看出, 有一部分成绩不是很理想的同学写下的熟悉同学非常多。因为此次问卷调查的时间是在学期末, 会出现部分学生为了通过该问卷拿到更高的平时分而故意多填的情况, 导致这个 friends_cnt 的统计值不是那么准确, 偏离了我们的本意。

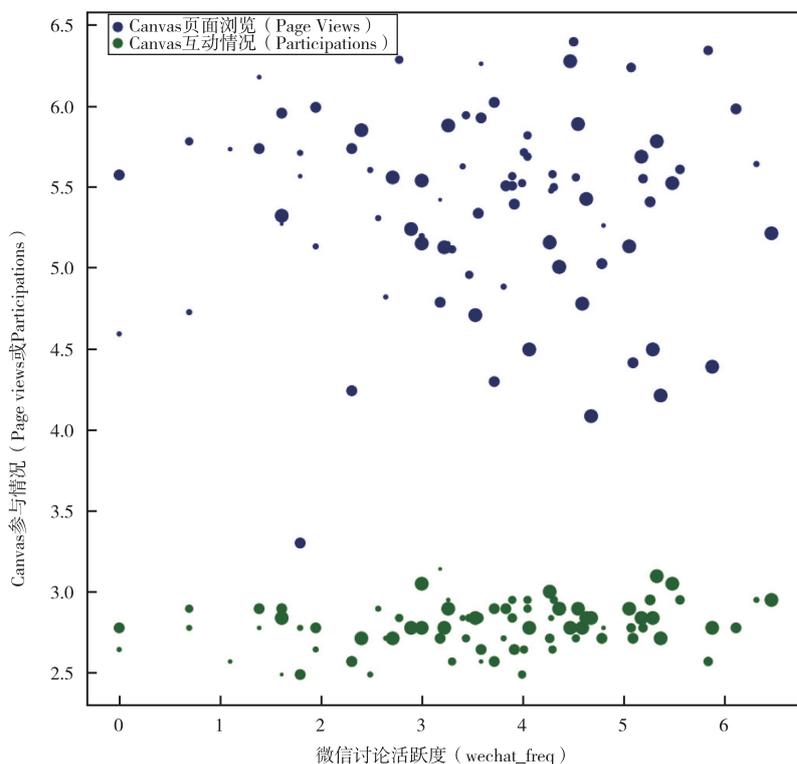


图 3 不同类型课程参与的关系

通过对成绩有正向影响的特征也能得到一些启发, 比如学生的竞赛基础越好, 则成绩越高; 期末模拟考成绩越高, 则期末成绩越高, 这都是符合常理的。此外可以观察到, 学生的平时表现、投入程度、互动情况、社交关系、自我评价都对成绩有显著影响。

(1) 平时表现: 模拟考分数高的学生期末分数更高, 作业按时提交并且作业分数高的学生期末成绩更高, 因为模拟考和作业都可以看作一种阶段性考核, 体现了学生对课程的掌握情况, 只是模拟考临近期末考试, 因此和期末成绩的正相关更明显。

(2) 投入程度: 关于投入程度有个很有趣的现象, 每周投入时间以及做过的练习题数目其实对成绩影响并不显著, 说明存在一些平时很努力、但是作业和考试分数都不是很理想的学生, 而一门本科生课程需要着重关注的, 就是如何帮助这部分学生学得更快更好。

(3) 互动情况: 学生的互动情况包括 Canvas 互动、线下答疑和微信群讨论, 由于课程讨论主要在微信群进行, Canvas 用来提交作业、下载课件和回看视频, 因此 Canvas 互动次数的意义不是很大, 线下答疑并没有显式出现在特征里, 所以主要关注微信互动情况, 实验结果也确实显示参与讨论较多的学生, 成绩可能更好。

(4) 社交关系: 学生的好友的平均分数越高, 学生的分数也越高, 侧面印证了“近朱者赤近墨者黑”这句话, 说明朋友之间一起学习, 或许学习效果比较好。而朋友数目这个特征, 上文已经讨论过, 成绩差学生会受到激励尽可能写多一点 (为了补救平时分), 因此体现出负向关系。

(5) 自我评价：通过问卷收集到了学生的学习成果自我评价，实验结果表明，认为自己学到了很多知识的学生倾向于拿到更高分数。可见，除了平时成绩（作业 + 模拟考）以外，学生的自我评价对课程评估与考核同样具有参考价值。

此外，学生的年级、性别、授课方式（线上回看、看直播、去教室听课）对成绩的影响并不显著。但是实际上，来自线下班级的学生成绩普遍更好，13班期末平均通过题目数为 2.74，中位数为 2.0；而 12班学生平均通过题目数为 4.20，中位数为 5.0，差距是非常大的。下一部分将具体分析产生上述现象的原因。

4 线上线下成绩差距对比分析

这一部分主要分析为什么线上线下平行授课的两个班级，成绩差距会如此明显。由于学校允许线下班级同学可以不来教室而是选择看直播或录像，线上班级的同学如果已经返校也可以到现场听课，所以两个班级的区别可能不完全在于授课方式，而是由于线下班级的同学都在校园里，可以面对面参与课程讨论和交流，而线上的同学可能在家里进行线上听课和讨论。由此进一步深入地挖掘两个班级的成绩差距的来源。

之前提到过，线下授课班级学生的成绩普遍更好，13班期末平均通过题目数为 2.74，中位数为 2.0，标准差为 1.76；而 12班学生平均通过题目数为 4.20，中位数为 5.0，标准差为 1.65，差距是非常大的，而且波动程度也差不多。排除掉有竞赛基础的学生（12班 9个、13班 2个）之后，13班期末平均通过题目数为 2.45，中位数为 2.0，标准差为 1.64；而 12班学生平均通过题目数为 4.10，中位数为 4.50，标准差为 1.66，差距仍然十分明显，可见这种成绩差距不是因为有竞赛基础学生较多，而是没有竞赛基础的大多数学生普遍存在。

下面，从上一节提出的影响学生成绩的 5 个因素出发，对比分析两个班级成绩差距的根本原因。

(1) 平时表现：从统计数据中可以看出，虽然作业平均分是 13班（58.55）略高于 12班（57.84），然而 13班的作业分数标准差为 11.72 要大于 12班的 6.52，13班最差的一位同学只拿到了 4 分。在作业按时提交率上，情况也类似，与线下授课的 12班相比，13班的作业按时提交率标准差为 14.35（12班为 5.49），最差的一位学生按时提交率只有一半（12班为 76.47）。而在模拟考成绩中，两班的差距进一步拉大，12班在模拟考中有 75% 的同学都做出了三道题及以上，而 13班超过 1.5 道题的只有一半，模拟考的成绩差距和期末考非常一致。

(2) 投入程度：这里只看 weekly_hours 和 oj_ACs 这两个特征，weekly_hours 的 13班均值为 3.23，小于 12班的 3.35，不过 13班的标准差 1.33 要大于 12班的 0.85，说明 13班的学生投入时间差距较大；而在 oj_ACs 这个特征上，13班均值 1.93 要略大于 12班的 1.82。

(3) 互动情况：13班在 Canvas 上的 Page View 平均为 438.83，明显高于 12班的 354.43，这是符合预期的，因为 13班同学不能线下参与课程，只能访问 Canvas。但是 13班参与微信讨论的平均次数仅为 20.83（12班为 73.95），并且标准差也很小，说明两班对比整体情况是：13班同学不太活跃，参与课程相关讨论较少。挑选出两个班级期末通过题数超过 3 个的同学（超过 3 个也

就是优秀), 12 班这些同学的微信讨论参与平均数是 79.49, 接近 12 班平均水平, 而 13 班通过题目超过 3 个的同学参与微信讨论的平均数是 41.29, 远远超过 13 班平均水平, 可见课程参与程度对线下授课班级的影响很大。

(4) 社交关系: 13 班学生朋友的平均分为 2.05, 显著低于 12 班的 3.65, 这和 13 班学生整体成绩较低有关, 同时也说明两班学生虽然课程进度、作业和考核都完全一致, 但是两个班级同学并没有什么交流。

(5) 自我评价: 从数据来看, 13 班同学的自我评价低于 12 班, 如对于“认为自己在本门课程中学到了很多知识”这一问题, 选择“非常同意”、“同意”和“一般”的学生, 在 13 班分别占 64.29%、28.57%、7.14%, 而 12 班占比为 73.58%、25.47%、0.94%, 说明 13 班认为通过课程学到很多东西的认可较 12 班稍差。

因此, 可以初步认为, 在平时表现、投入程度、互动情况、社交关系和自我评价方面, 两班之间都有一些差距。而其中平时表现、社交关系和自我评价其实都是结果性质的指标, 并不是两班的本质区别。也就是说, 两班的主要区别体现在课程投入程度和互动情况上, 13 班学生互动情况差, 投入时间少, 学习效果整体较差, 而且学生之间差距大。

因此, 对于线上教学来说, 是否通过视频授课其实并不是成绩的关键影响因素, 而是线上授课目前无法提供学生需要的氛围感和参与感, 导致学生的学习主动性较差。12 班几乎所有同学都曾至少参与过一次线下听课 (学校不要求学生必须每次线下参与), 并且在校园里可以线下讨论和提问, 微信群也较为活跃。而 13 班学生可能是待在家里通过看直播或者视频回看来参与课程 (实际上只有 60% 的人看过直播), 他们没有身处校园, 学习氛围与在校学生很不一样, 而且 13 班微信讨论也少, 学生很难有和伙伴共同学习的感觉, 遇到问题也较少交流讨论。面对线上教学目前存在的这类问题, 可以通过丰富的工具来营造学习氛围感、提升学生参与度, 这应该成为线上课程今后改进的一个发力点。

5 学生成绩预测

由于前面已经分析了不同因素对学生成绩的影响, 以及线上线下两班成绩差别的原因, 这一部分使用回归模型, 尝试对学生成绩做出预测^[15]。由于多元线性回归模型较为简单, 因此还尝试了机器学习算法中的随机森林^[16]和 Xgboost^[17]这两种方法。

随机森林和 Xgboost 都是基于决策树的算法。决策树是机器学习中的一个预测模型, 反映了对象属性与对象值之间的一种映射关系。在决策树中, 每个节点对应一个对象, 每个分支则对应一个属性值, 每个叶节点则对应从根节点到该叶节点所经历的路径所表示的对象的值。除此之外, 决策树只有唯一的一个输出。

随机森林是一个包含多个决策树的分类器, 并且其输出的类别是由个别树输出的类别的众数而定。它是决策树为基础的一种更高级的算法。与决策树一样, 随机森林既可以用于回归, 也可以用于分类。从名字中可以看出, 随机森林是用随机的方式构建的一个森林, 而这个森林是由很多相互不关联的决策树组成的。事实上, 随机森林从本质上属于集成学习, 是机器学习的一个

很重要的分支。集成学习通过建立几个模型的组合来解决单一预测问题。其工作原理是生成多个分类器（模型），各自独立地学习和作出预测。这些预测最后结合成单预测，因此优于任何一个单分类做出的预测。

梯度提升决策树（Gradient Boosting Decision Tree, GBDT）是一种基于 boosting 集成思想的加法模型。训练时采用前向分布算法进行贪婪地学习，每次迭代都学习一棵 CART 树来拟合之前 $t-1$ 棵树的预测结果与训练样本真实值的残差。Xgboost 则是梯度提升方法在决策树上的一种实现，具有较为完善的工程支持。它是一个优化的分布式梯度增强库，旨在实现高效、灵活和便携。它在 Gradient Boosting 框架下实现机器学习算法。Xgboost 提供了并行树提升（也称为 GBDT、GBM），可以快速准确地解决许多数据科学问题。相同的代码在主要的分布式环境（Hadoop、SGE、MPI）上运行，并且可以解决超过数十亿个样例的问题。

这里将全部学生按 80% 和 20% 的比例划分为训练集和测试集，所有模型都在训练集上拟合参数，在测试集上测试效果。在评测指标上，选择 mean absolute error (MAE)，也就是预测值和真实值的差的绝对值，MAE 值越小，说明回归结果越接近真实值，模型效果越好。其中，随机森林的参数为：estimators 为 60，每棵树的 features 是 8；Xgboost 的参数为：学习率 0.3，estimators 为 16，其余均用默认值^[18]。

表 3 展示了三种模型对学生成绩预测的效果，可见随机森林要优于线性回归，因为它是一个机器学习模型，更为复杂，并且利用了模型集成的思想来降低结果的方差。然而更复杂的 Xgboost 模型并没有取得更好的效果，原因是其参数更多、比较难调试，而且更复杂的模型也意味着更容易过拟合。在训练集上观察到的拟合效果确实是 Xgboost > 随机森林 > OLS。

这里的 MAE 值都在 1 以下，而期末试题是 6 道题，也就是说分数在 0 ~ 6 的范围内，表 3 模型在测试集上对学生成绩的预测误差绝对值平均小于 1，能够在实际教学中为教师提供一些指导性意见，或者对部分表现欠佳的学生及时发出预警。

表 3 学生成绩预测效果

方法	测试集 MAE 值
OLS	0.960
Random forest	0.931
Xgboost	0.987

6 总结与展望

通过对 2020 年秋季编程入门课程“计算概论 (B)”的学生行为及表现的分析，得到以下结论及教学建议。

6.1 学习效果评估方面的教学建议

(1) 学生的作业成绩和模拟考试成绩与期末成绩直接正相关，所以重视学习过程中的阶段性

评估十分重要, 在教学实验中弱化期末单次考试的影响或许是可行的。

(2) 在本课程中, 有编程竞赛基础的学生非常占优势, 两个班有竞赛基础的总共 11 个学生, 期末 6 题平均通过 5.5 题, 或许可以让这部分学生直接免修或者安排到提高班, 不挤占其他学生的优秀率。

6.2 优化线上教学设计方面的教学建议

(1) 来自线下授课班级的学生成绩普遍更好, 包括作业、模拟考和期末成绩都表现出这样的特性, 并且线上授课的同学认为自己学到了知识的比例偏低, 说明线上授课在主客观方面的教学效果较差, 建议教师和助教对线上参与的学生多一些关注, 如定期访谈、询问学习体验, 并对遇到困难的学生进行额外的辅导^[19]。

(2) 并不是平均每周投入时间越多成绩就越好, 说明正确的学习方法很重要。除了学生总结自身的学习方法之外, 提升学生的学习效率也应该是课程设计要考虑的内容。关于这部分, 本门课程已经做了一些初步的尝试, 包括鼓励学生整理学习笔记本和心得体会等, 按学生意愿分享给下一届的学生, 期望每位学生能从这些经验中找到适合自己的学习方法。

(3) 线上、线下两班学生成绩差异大的根源不是竞赛基础差距, 也不能认为是不同授课方式的区别, 线上授课缺少的是足够的互动和学习氛围, 而只要互动够, 直播 / 录播这种授课方式本身或许没有问题, 毕竟 13 班的学生明显表现出参与微信讨论越多, 则成绩越高。传统课堂中, 学生的课上互动主要靠老师带动, 而课下互动则主要依赖于学生空间上处于相近的位置, 并且彼此较为熟悉。对于线上教学来说, 如何设计和鼓励学生参与课上和课下讨论交流, 是重中之重。关于这部分, 如果是线上线下平行班的情况, 可以让两班同学一起讨论, 如: 加入相同的讨论群组, 或许线下班级的气氛能带动远程参与的学生; 在课程初期在线上助教课上安排“破冰”活动, 在课程中后期安排学生分享作业心得等; 考虑以平时课堂表现分作为奖励^[20]。

(4) 线上教学平台会收集非常多的学生行为数据, 这些数据蕴含丰富的信息, 可以根据这些数据构建学生成绩评测模型, 用于阶段性地向教师示警存在落后的同学, 也有助于教师对学生整体掌握水平有初步了解, 从而适当调整课程节奏^[21]。

6.3 后续优化方向

本研究存在的潜在的优化方向, 主要列举以下几个:

(1) 更细粒度的特征。目前模型利用的特征只有二十几个, 有很多很有启发性的特征还没有利用, 比如完成作业的时间、作业发布后首次浏览用时、微信提问次数和回答次数、微信发言字数、学生社交关系图中节点的中心性程度等。

(2) 进行自动特征筛选。由于使用的特征来源包括 Canvas 平台、课程微信群、多次问卷调查和学生人际关系, 一旦添加更细粒度的特征, 特征维度会变得更高。现在只利用了其中的一部分特征, 后续考虑充分利用更多特征的信息, 再通过 PCA、逐步回归、随机森林等方法对特征降维, 提取关键信息。

(3) 考虑采用 few-shot learning 方法, 也就是在数据量较少的情况下, 如何得到更准确的模型预测结果。目前一门课程只有一百多个学生, 数据量显然不够。这种情况下, 可以用 few-shot learning 的方法, 尽量优化模型在这仅有的样本上学到的信息, 还可以想办法扩充相关数据, 采取预训练的方式, 起到同样的效果。

【参考文献】

- [1] 教育部:利用网络平台,“停课不停学”[EB/OL]. [2020-08-22].http://www.moe.gov.cn/jyb_xwfb/gzdt_gzdt/s5987/202001/t20200129_416993.html.
- [2] EDUCAUSE.2020 EDUCAUSE Horizon Report:Teaching and Learning Edition [EB/OL]. [2020-08-02].<https://library.educause.edu/resources/2020/3/2020-educause-horizon-report-teaching-and-learning-edition>.
- [3] 王志军,陈丽.国际远程教育教学交互理论研究脉络及新进展[J].开放教育研究,2015,21(2):30-39.
- [4] International Council for Open and Distance Education.New Report on Ethics in Learning Analytics [EB/OL]. [2020-08-23].<https://www.icde.org/icde-news/new-report-on-ethics-in-learning-analytics>.
- [5] ADL.Experience API Version 1.0.1 [EB/OL]. [2016-02-01].<https://github.com/adlnet/xAPI-Spec/blob/master/xAPI.md>.
- [6] 李晓明.与在线学习一起成长[J].计算机教育,2020(6):8-11.
- [7] 陈轶.疫情期间基于OBE理念在线PBL教学的探索和实践[J].计算机教育,2020(6):25-29.
- [8] 徐国艳.基于学习成效金字塔理论的远程混合式教学设计与实践[J].计算机教育,2020(6):17-20.
- [9] Spady W G.Outcome-Based Education: Critical Issues and Answers [M].Arlington:American Association of School Administrators,1994.
- [10] Dolmans D H,De Grave W,Wolfhagen I H,et al. Problem-based learning: future challenges for educational practice and research [J].Medical Education,2005,39:732-741.
- [11] 姜艳玲,徐彤.学习成效金字塔理论在翻转课堂中的应用与实践[J].中国电化教育,2014(7):133-138.
- [12] 黄德群.云服务架构下的Canvas开源学习管理系统研究[J].中国远程教育,2013(7):64-70.
- [13] Torre M V, Tan E, Hauff C. edX log data analysis made easy: introducing ELAT: An open-source, privacy-aware and browser-based edX log data analysis tool [C].Proceedings of the Tenth International Conference on Learning Analytics & Knowledge. 2020: 502-511.
- [14] Viberg O, Khalil M, Baars M. Self-regulated learning and learning analytics in online learning environments: A review of empirical research [C].Proceedings of the tenth international conference on learning analytics & knowledge. 2020: 524-533.
- [15] Van Goidsenhoven S, Bogdanova D, Deeva G, et al. Predicting student success in a blended learning environment [C].Proceedings of the Tenth International Conference on Learning Analytics & Knowledge. 2020: 17-25.
- [16] Ho T K. Random decision forests [C].Proceedings of 3rd international conference on document analysis and recognition. IEEE, 1995, 1: 278-282.
- [17] Chen T, He T, Benesty M, et al. Xgboost: extreme gradient boosting [J]. R package version 0.4-2, 2015, 1(4): 1-4.
- [18] 吴砥,饶景阳,吴晨.教育大数据标准体系研究[J].开放教育研究,2020(4):75-82.
- [19] 张福涛.《2020地平线报告:教与学版》的解读及思考——疫情之下高等教育面临的挑战与变革[EB/OL]. [2020-08-22].http://www.360doc.com/content/20/0524/18/30898787_914290774.shtml.
- [20] Liu Z, Yan H, Chen C, et al. Evaluation of Learning Effect Based on Online Data [C]//China Conference on Information Retrieval. Springer, Cham, 2021: 141-154.
- [21] 向开成,王浩,唐莉,等.疫情期间在线教学效果的影响因素与对策研究——基于4350份中国广西大学生问卷调查分析[J].教学方法创新与实践,2020(8):37-40.

Online Data Analysis and Mining for Canvas Learning Management System

Yan Hongfei^{1,2} Hu Yang¹ Wang Yankai¹

(1. School of Computer Science, Peking University, Beijing 100871, China;
2. Center for Data Science, Peking University, National Engineering Laboratory for
Big Data Analysis and Applications, Beijing 100871, China)

Abstract: [**Purpose/significance**] The COVID-19 pandemic has brought the prospect of online teaching and online and offline mixed teaching. However, people still generally believe that students underperform with online teaching compared to offline classroom teaching. This article aims to analyze the fundamental differences between these two teaching methods, and to explore how to fully exploit the potentials of online teaching, so that students may have the best learning experience. [**Method/process**] This article chooses online and offline parallel classes of the undergraduate course *Introduction to Computing(B)*, analyze the key influencing factors of students' academic performance and make predictions about students' final scores, with the data provided by learning management system and collected through questionnaire. [**Result/conclusion**] This paper believes that the key factors affecting students' grades include their daily performances, levels of engagement, course interactions, social relationships and self-assessments. The differences between the two teaching methods is mainly reflected in the degree of course commitment and interaction. Finally, combined with the actual situation, teaching suggestions are put forward from two aspects: learning effect evaluation and optimization of online teaching design.

Keywords: Regression analysis; Learning management system; Online learning

(本文责编: 孙龙慧)