文章类型：论文|刊号（ISSN）：2630-5178 /（中图刊号）：380GL019

基于随机森林的工程管理专业课程设置对毕业表现影响的实证分析

丁泓翔* 葛立杰 徐玲玲
河北建筑工程学院经济管理学院
DOI:10.12238/mef.v8i11.14800

[摘 要] 本文通过构建随机森林分类模型对河北某高校240名工程管理专业毕业本科生四学年的成绩数据进行了分析,阐述了通识教育类课程、学科教育基础课程、专业教育课程(含实习)的课程设置对学生毕业表现(毕业设计(论文)成绩)的影响重要性差异。结果表明,随机森林模型具有较高的预测准确性,学科教育基础课程对毕业表现的影响最为显著。本研究为优化课程体系、提升工程管理专业人才培养质量提供了科学依据。
[关键词] 随机森林；工程管理；教学；毕业表现；课程设置
中图分类号：G424.1 文献标识码：A

An Empirical Analysis of the Influence of Engineering Management Curriculum on Graduation Performance Based on Random Forest

Hongxiang Ding* Lijie Ge Lingling Xu
Hebei University of Architecture, School of Economics and Management

[Abstract] This article analyzes the academic performance data of 240 undergraduate students majoring in engineering management at a university in Hebei over four academic years via a random forest classification model. It reveals the differences of the influencing importance across the general education courses, basic disciplinary education courses, and professional education courses (including internships) in influencing students' graduation performance (graduation project/thesis grades). The results show that the random forest model has a high prediction accuracy, and the basic courses of subject education have the most significant impact on graduation performance. This study provides a scientific basis for optimizing the curriculum system and improving the quality of engineering management professionals.
[Key words] Random forest; Engineering management; Education; Graduation Performance; Curriculum.

随着我国建筑行业发展的积极转型,对新时代背景下应对新问题的管理人才的需求正在日益增加。工程管理专业课程作为培养工程管理人才的核心环节,直接关系着人才培养的质量,对高校深化教学改革具有重要指导意义。但工程管理专业的课程设置和教学体系具有一定的滞后性。为避免课程体系与行业、社会需求之间产生脱节,需对课程设置进行及时调整[1-2]。近年来,随着教育技术的发展和高校学生人数的不断增加,高校数据库中已经积累了大量的学生课程成绩数据[3-4]。同时机器学习技术的快速发展又为教育数据的挖掘提供了新的思路。目前,大部分高校只是对以上学生数据进行了浅层的统计、查询,而没有使用教育数据挖掘技术对学生课程成绩数据中深层的价值信息进行提取和研究,可见通过探索一种能够对学生专业课程成绩数据进行科学评价的分析方法,进而对教学工作进行指导以满足行业的快速变化带来的严峻考验就显得尤为紧迫[1-5]。

基于此,本研究将使用河北某高校工程管理专业毕业生大学四年各课程成绩和最终毕业表现数据构建随机森林分类模型,检验模型可靠性的同时分析工程管理专业课程设置与学生毕业表现之间的内在联系,识别出影响学生毕业表现的关键课程因素,并根据各课程对毕业表现的影响重要性差异提出有针对性的课程建设建议。

1 随机森林的应用及原理

随机森林算法(Random Forest)最早由Leo Breiman提出,随机森林是一种组成式的有监督学习方法,本质是对多个决策树决策结果进行汇总的多决策树模型[6]。相比于传统预测算法,

文章类型：论文|刊号（ISSN）：2630-5178 /（中图刊号）：380GL019

随机森林具有参数调节简单、独立训练速度快、可进行变量对分类的相对重要性评估、可避免“过拟合”等优点。另外,随机森林通过装袋法(Bagging)集成学习构建多个决策树,可以有效降低模型的方差,提高预测精度。

随机森林已经被广泛的应用于学生成绩分析的研究之中。如Hoz等人使用随机森林模型对哥伦比亚2018年256所大学的工学课程考试数据进行了建模分析。结果显示随机森林具有良好的预测精度,且部分课程(如英语、数学)对考试成绩具有显著的重要影响[5]。宋园等人选用某高校09级计算机学院379名学生课程成绩通过随机森林算法评价了各课程成绩对学生大三下学期专业学习成绩的影响。结果显示部分前期课程对后续具有显著影响[3]。根据以往研究不难发现,随机森林方法应用场景广泛、模型构建灵活。但以往应用随机森林模型的研究鲜有针对工程管理专业课程设置的深入分析。本研究将弥补这一研究空白。

随机森林构建决策树时会采用有放回抽样方法(Bootstrap)取样,从而确保每棵决策树既有相似性又有差异性。基于该取样方法,会存在部分从未被选中的样本(被称为袋外数据),故很多情况下随机森林不需要额外设置测试集数据,只需要用袋外数据来测试模型即可[6,7]。

随机森林实现步骤可概括如下：

(1)假设数据训练集中共有N个样本、M个变量,根据Bootstrap采样规则从训练集中随机有放回地抽取N个样本用于训练根节点,构建决策树。

(2)在每一个节点随机抽取m(m＜M)个变量,将其作为分割该节点的候选变量,因为每个节点都是随机抽取变量,故各个节点的候选变量大概率不相同。同时每一个节点处的候选变量数应一致。分别计算候选变量的Gini指数,Gini指数越小表示在这个集合中被分错的概率越小,根据最小准则选择指数最小变量,设定为节点分裂变量。其中Gini系数在节点t处的计算公式如下：

$$Gini(t)=\sum_{k=1}^{K}p_k(1-p_k)=1-\sum_{k=1}^{K}p_k^2$$

式中,K为分裂变量中分类选项的总数量,k为K中的某个类别, p_k 表示样本属于类别k的概率。

(3)完整生成决策树。

(4)重复(1)-(3)过程,生成大量相互独立且重要性相等的决策树;终端节点的所属类别由节点对应的众数类别(或平均数)决定。

(5)对于用于预测的新数据,用所有的树对其进行分类,其类别由多数决定原则(或平均数)生成。

2 工程管理专业课程分析模型构建

2.1数据来源及数据预处理

本研究选取河北某高校18级、19级经济管理学院工程管理专业240名毕业本科生四学年共13920条成绩数据作为研究数据。研究模型包含工程管理专业通识教育类课程、学科教育基础课程、专业教育课程(含实习)、毕业设计(论文)等课程。

数据中存在部分缺失、无效值等问题,会对模型的训练精度造成干扰,为保证数据和模型的质量需先对数据进行预处理。具体预处理方法如下：(1)删除所修课程差异较大的转专业学生数据。(2)合并、删减部分重复课程的数据。(3)删除因辍学、缺考、作弊而产生大量课程成绩缺失的学生数据。(4)针对少量课程出现缺失的数据采用年级该课程成绩平均值来替代。(5)删除部分选修课成绩数据。6)对学生成绩进行离散化处理,将每门课的成绩划分为五个等级,不及格([0,60])、及格([60,70])、中([70,80])、良([80,90])、优([90,100])。

2.2模型实现

本研究采用R语言软件对随机森林模型进行训练和构建。随机森林模型选取学生毕业表现作为最终决策节点,其他课程成绩作为根节点及内部节点变量。随机森林模型训练需要确定随机森林生成的决策树数量和节点分裂时随机选取的备选特征变量的数量m。通过模型试验发现本研究模型在决策树大于200时误差趋于稳定,故本研究生成的决策树数量选取为200。根据Leo Breiman的建议[6],节点分裂时随机选取的备选特征变量的数量可取总特征数量M的平方根,本文选取m为8。本研究随机森林模型为分类模型,由众多决策树投票确定。

3 结果分析

因为本研究采用的学生数据为毕业生数据,绝大部分学生都完成了不及格课程的补考或清考,课程成绩均采用补考后的最终成绩,故本研究数据中仅包含及格、中、良、优四个成绩级别的分类。本研究选取学生的毕业设计(论文)成绩来量化学生的毕业表现。

通过模型训练,本研究随机森林模型对工程管理专业毕业生毕业表现的预测准确率高达79.6%,模型精度较高。本研究还针对各个课程成绩对毕业表现的影响重要性进行了排序和比对。基于Gini系数最小原则,为获取最小Gini系数,需获得最大的Gini系数减少量(即最大的Mean Decrease Gini),故本研究以R语言提供的Mean Decrease Gini指标作为评价各特征变量重要性的指标,数值越大代表该特征变量的重要性越大。

图1为各工程管理专业课程对学生毕业表现影响重要性的排序前30的统计结果图。由图1可知学科教育基础课程占12门,其中西方经济学、高等数学B、工程力学、画法几何与建筑制图、工程测量更是位列前10。究其原因可能是因为工程管理专业是土木工程和管理学相融合的交叉学科,需要学生通过学科教育基础课程掌握必要的数学知识、经济学知识和工程知识,而以上扎实的知识学习基础又有利于学生深入理解毕业设计(论文)的工程内容并熟练地运用数学力学方法解决工程问题,进而取得更好的毕业表现。

在专业教育课程中工程估价、施工组织和工程造价管理三门课程同样位列最重要的课程前10名,可能是因为专业教育课程是针对工程管理专业知识的进一步细化,更为贴合学生毕业就业的专业方向的技能需求,部分学生毕业设计会选择工程招

投标文件编制和施工组织文件编制内容，与专业教育课程契合度较高。

实习实践类课程中有四门课程对学生毕业表现产生了较大影响，其中生产与管理实习课程对学生毕业表现影响最大，可见学生知识的学习不能仅仅停留在课堂，认真践行现场实践教学环节的学生更容易获得更好的毕业表现。

本研究的通识教育类课程重要性结果中马克思主义基本原理具有最大的影响重要性。可能是因为马克思主义基本原理课程揭示了社会发展的基本规律，为学生提供了辩证历史唯物主义的科学方法，提升了学生独立思考和分析问题的能力。英语类课程和体育类课程同样具有一定的重要性，可能是因为外语技能可以帮助学生更好的阅读外文文献，而体育类课程成绩则体现了学生的身体状态，更好的身体状态可以保证学生专注顺利地完成毕业设计（论文）工作。

综上所述，一个合格的毕业表现出色的工程管理专业学生需要掌握一定的经济学理论、数学基础、工程类专业知识、管理学理论、外语技能，并具备一定的实习实践经历，同时还需具有一定的思政理论和健康的体魄。

针对以上结果分析，本研究提出以下一些针对工程管理专业课程建设的建议：

（1）强化学科教育基础课程教学，应该涵盖工程知识、经济学知识、数学知识的基础课程，针对骨干课程进行教学融合，增设跨学科衔接课程。

（2）深化专业教育课程，以就业需求为导向进行课程设计，确保选题与实际工程问题相结合。

（3）紧抓实践实习质量，合理优化学生实践实习内容。构建阶梯式实践体系，加强校企合作平台建设。

（4）注重学生通识教育类课程的学习情况，引导学生用辩证唯物主义分析工程问题。拓展英语、体育课程教学功能。

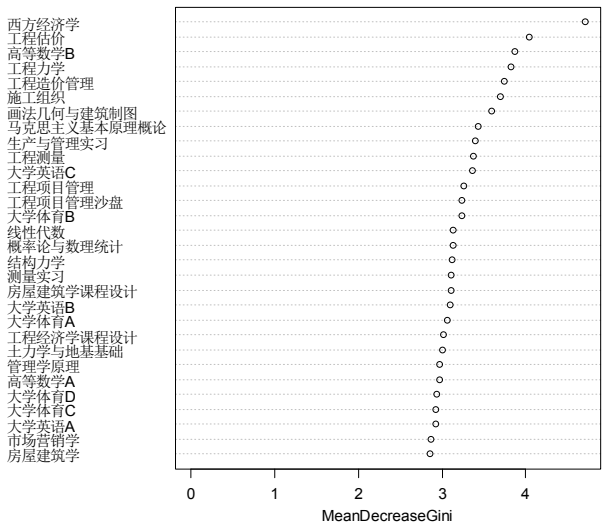

图1 课程重要性排序

4 结论

本研究选用河北某高校工程管理专业毕业本科生的成绩数据构建了随机森林分类模型，探讨了各课程成绩与毕业表现之间的内在联系。并针对结果提出了工程管理专业课程建设的建议。本研究的研究结论如下：

（1）随机森林模型能够有效识别影响学生毕业表现的关键课程因素。

（2）影响学生毕业表现的课程重要性呈现一定的层级差异，学科教育基础课程起到核心支撑作用。这些课程显著提升毕业设计（论文）的理论深度和技术应用水平。

（3）部分专业教育课程因与毕业设计内容和就业需求高度契合，成为了提升毕业表现的关键课程。

（4）生产与管理实习等实践实习课程通过现场经验积累，显著增强学生对理论知识的转化能力。

（5）大学通识教育课程奠定了学生的辩证思维、外文阅读技能、强健体魄等综合素养，可间接支持学术与实践任务的完成。

（6）需进一步深化、优化工程管理专业课程的课程设置以培养与行业需求相符的高素质工程管理专业人才。

[基金项目]

河北省高等教育教学改革研究与实践项目：基于随机森林的建筑产业转型升级背景下工程管理类专业实践教学体系改革研究，项目编号：2023GJJG335。

[参考文献]

[1]余弦，周谊芬.大数据背景下基于随机森林算法的高校学生成绩预警研究[J].江苏科技信息，2020，37(20)：50-53.

[2]吴兴惠，周玉萍，邢海花，张大帅.基于机器学习算法的高校学生成绩评价模型的构建[J].数码精品世界，2022，(7)，52-54.

[3]宋园，朱丽琴，程泽凯.基于随机森林的学生成绩评价研究[J].齐齐哈尔大学学报(自然科学版)，2017，33(06)：1-5+10.

[4]王萱，刘勇，朱常香，高姗，陈海亮.生物技术专业学生平时成绩与考研成功率间的随机森林预测分析[J].科教文汇，2023，(24)，99-102.

[5]Hoz，E.D.L.，Zuluaga，R.，& Mendoza，A.Assessing and classification of academic efficiency in engineering teaching programs[J].Journal on Efficiency and Responsibility in Education and Science，2021，14(1)，41-52.

[6]Breiman，L.Random forests[J].Machine learning，2001，45(1)：5-32.

[7]Breiman，L.Bagging predictors[J].Machine learning，1996，24(2)：123-140.

作者简介：

丁泓翔(1991--)，男，河北省张家口市人，博士，讲师，研究方向为管理学、心理学、交通规划等交叉学科研究。

葛立杰(1976--)，男，河北省保定市人，硕士，副教授，研究方向为工程管理数字化。

徐玲玲(1978--)，女，河北省石家庄市人，硕士，副教授，研究方向为工程管理、施工组织、建筑低碳。