Advances in Education 教育进展, 2024, 14(5), 571-578
Published Online May 2024 in Hans. https://www.hanspub.org/journal/ae
https://doi.org/10.12677/ae.2024.145735

Hans汉斯

产教融合目标下数据科学与分析教学案例设计

王　昕

北京信息科技大学理学院，北京

收稿日期：2024年4月16日；录用日期：2024年5月14日；发布日期：2024年5月21日

摘　要

当前，我国高校要创新人才培养模式，建立健全多层次、多类型的大数据人才培养体系，重点培养具有统计分析、计算机技术、经济管理等多学科知识的跨界复合型人才。基于上述，针对大数据人才培养的核心课程，结合实际应用背景，设计符合学生理解能力和应用能力的教学案例，使用较大规模数据集和机器学习算法，分析影响银行客户流失的各类特征及其重要性排序，并用平均绝对误差等指标评估模型性能，分析提升模型性能的方法，力求提高理论课与实践教学的交互水平，致力于培养具有较强实践能力和创新意识的高素质应用型创新人才具有重要意义。

关键词

数据科学，机器学习，案例设计，银行客户流失

The Design of the Teaching Case for Data Science and Analysis under Production and Education Integration Goal

Xin Wang

School of Applied Science, Beijing Information Science and Technology University, Beijing

Received: Apr. 16th, 2024; accepted: May 14th, 2024; published: May 21st, 2024

Abstract

At present, colleges and universities in China need to innovate the talent cultivation mode, establish and improve a multi-level and multi-type big data talent cultivation system, and focus on cultivating cross-border compound talents with multi-disciplinary knowledge such as statistical analysis, computer technology and economic management. Based on the above, in view of the core curriculum of big data talent cultivation, combined with the actual application background, the teach-

文章引用: 王昕. 产教融合目标下数据科学与分析教学案例设计[J]. 教育进展, 2024, 14(5): 571-578.
DOI: 10.12677/ae.2024.145735

ing case that meet the students' understanding ability and application ability are designed. Using large-scale data sets and machine learning algorithms, various characteristics and importance rankings that affect bank customer churn are analyzed, and the average absolute error and other indicators are used to evaluate the performance of the model, and the methods to improve the performance of the model are analyzed. It is of great significance to strive to improve the interaction level between theoretical and practical teaching, and to cultivate high-quality applied innovative talents with strong practical ability and innovative consciousness.

Keywords

Data Science, Machine Learning, Case Design, Bank Customer Churn

1. 引言

“数字经济”和“数字中国”自 2017 年首次被写入《政府工作报告》以来，已连续多年被重点提及。中央指出，建设数字中国是数字时代推进中国式现代化的重要引擎，是构筑国家竞争新优势的有力支撑。加快数字中国建设，对全面建设社会主义现代化国家、全面推进中华民族伟大复兴具有重要意义和深远影响[1]。为不断加强大数据专业人才培养力度，至 2023 年全国有 757 所高校成功申报“数据科学与大数据技术”本科专业[2]。

课题组对大数据专业的核心课程数据科学与分析教学案例进行改革，以银行抵御风险的社会实际应用为背景，以数据预处理、描述统计、模型评估、算法提升等为手段，设计出不同层次、难度的课程案例，立足于产教融合，把专业培养目标融入课程内容，有效提升教学效果，对有志于从事数据分析与机器学习的学生在未来求职有实质性帮助。

2. 数据科学与分析案例教学设计理念

数据科学涉及多个学科领域，如统计学、计算机科学、数学等，同时，由于数据科学领域的快速发展，在教学设计中，不仅要注重理论知识的传授，确保学生具备扎实的理论基础，更应该注重跨学科知识的融合，具有一定的实用背景。因此，课题组对数据科学与分析教学案例进行设计时，着意以真实世界中的问题或挑战作为教学起点，引入行业或领域内的经典案例，如来自于如金融、医疗、电商、社交媒体等不同行业的案例，这些案例直接来源于现实世界、具有实际应用价值，且具有行业代表性，能够反映出该领域的典型问题和挑战。案例通常具有足够的数据量、复杂度和挑战性，能够激发学生的探索欲望，并为学生提供深入分析和学习的机会。此时学生面对的教学内容不仅仅是数据技术问题，更是与实际业务紧密相连的现实问题，通过数据科学技术来解决问题，可以更好地理解数据科学在实际操作中的作用，提升学生的问题解决能力和跨学科思考能力。

同时，设计教学案例时应确保该案例有明确对应的知识点，并注意教学案例内容的完整性和系统性。因此，案例中应包含数据预处理、描述性统计、特征选择、模型选择、模型训练、模型评估等各个步骤，通过细化每一个案例的实践环节，更加清晰地传达数据分析的核心知识和技能，帮助学生建立起完整的知识框架。此外，还需要充分考虑到学生的参与度和互动性，通过积极参与和互动，学生不再是被动的

知识接受者，而是成为主动的探索者，他们需要亲自动手进行操作和分析，这不仅能够加深对理论知识的理解，还能锻炼实际操作能力。教师协助学生理解案例场景，确定分析步骤，鼓励学生对案例进行多角度的分析和思考并提出自己的见解和解决方案。学生自主分组协作、程序编制、报告撰写、课堂展示等具体方式，每人以不同的分工角色积极参与到案例的分析和解决过程中，并得到教师及时的反馈和建议。通过案例式教学设计，学生可以接触到真实的数据分析场景和问题，亲身参与和互动，感受到学习的乐趣和价值，从而激发他们的创新思维和解决问题的能力，为终身学习打下坚实的基础。

最后，灵活多样的课堂教学评估方法对于提高案例教学效果至关重要，在该门课程的教学过程中，课程团队采用了线上和线下相结合的评估方式。首先，在课前，通过线上教学平台发送教学视频和相关案例视频或资料，设置相关课前测，教师通过线上教学平台收集数据，并进行分析，了解学生对即将学习的内容的掌握情况，发现学生可能存在的疑难点和误区，有助于调整教学策略；课堂内采用口头问答与讨论的方式，鼓励学生积极思考和表达观点，教师根据学生的回答内容和逻辑性进行即时评价，给予反馈和剖析；在课后，指导学生分组共同完成任务，并在下次课进行课堂展示，教师对小组的合作过程、实验报告的成果展示以及学生的实践能力和互动表现给予点评，并鼓励学生之间互相评价案例报告，通过互评和互助的方式，发现彼此的优点和不足，相互学习和提高。

3. 数据科学与分析“产教融合”目标教学案例

以金融背景的教学案例为例，银行是一个国家资金储存和周转的重要载体，在互联网发展迅猛的今天，线上支付方式对传统银行的影响很大，此时银行需要从维护客户层面积极思考应对策略，增强应对经济周期波动的能力。因此对银行客户流失的影响因素分析及预测非常重要[3]。银行客户流失是一个典型的分类问题，这一案例为学生提供了一个实际且具挑战性的场景，学生可以通过收集和分析客户的交易记录、个人信息、服务使用情况等数据，建立预测模型来识别哪些客户有流失的风险，同时更好地掌握数据分析的基本流程和技能。

本文数据是从网站 DataFountain 中获取[4]，里面共包含 10,000 个样本，共 14 个变量，包括客户编号、客户身份、姓氏、信用分数、地理位置、性别、年龄、任期、帐户余额、产品数量、是否有卡、是否为活跃会员、预估薪资以及是否流失等。本文旨在研究银行客户流失的相关问题，所以选择“流失”作为决策变量，其余变量则作为影响特征，具体的案例处理步骤如下。

3.1. 数据预处理

原始数据中存在一些特征与本研究无关，初步处理数据集时需要删除这些特征，如“行数”、“客户 ID”以及“姓氏”这三个特征作为标识性特征，是为了保证数据的有效性。在研究客户流失的影响因素时，这些特征属于无关特征，删除上述无关特征之后，保留的有效特征及其相关说明如表 1 所示。

由于数据库所用为编码数据，在分类处理时把“性别”“是否有信用卡”“是否为活跃客户”和“是否流失”这几个变量按需由字符型数据转化为数值型数据，对应编码数据如表 2 所示。

3.2. 描述性统计

对各个特征进行描述性分析，以“性别”特征为例，做该特征与“流失”的交叉列联表，如表 3 所示，样本集中共有男性客户 5457 个，占比 54.57%，女性客户 4543 个，占比 45.43%。在女性客户中 74.9%未流失，25.1%流失，未流失与流失的客户数量比例约为 3:1，流失比例稍高；男性客户中 83.5%未流失，16.5%流失，未流失与流失客户比例约为 5:1。女性的流失比率比男性高，说明女性客户更不稳定。

图 1 为“性别”特征在“流失”与“未流失”两种客户分类下的频数簇状条形图。可以直观地看出，

男性在基数略大于女性的基础上，未流失客户中男性远远多于女性，更加印证了女性客户的流失倾向比男性更严重，在银行的客户流失检测中要着重注意女性客户。

Table 1. Table of variable description
表 1. 变量描述表

特征名称	特征缩写	变量类型	特征说明
Credit Score	Score	数值型	信用度
Gender	Gender	字符型	性别
Age	Age	数值型	年龄
Tenure	Tenure	数值型	任期
Balance	Balance	数值型	帐户余额
Numof Products	Num	数值型	购买本行产品数量
Has CrCard	Card	字符型	有信用卡
IsActive Member	Act	字符型	活跃客户
Estimated Salary	Salary	数值型	预估薪资
Exited	Exited	字符型	流失

Table 2. Table of feature coding
表 2. 特征编码表

特征缩写 / 编码	Gender	Card	Act	Exited
0	女	无	否	否
1	男	有	是	是

Table 3. Exited* Gender cross table
表 3. Exited* Gender 交叉列联表

			Gender		合计
			0 (F)	1 (M)	
Exited	0	计数	3404	4559	7963
		Gender 中的%	74.9%	83.5%	79.6%
	1	计数	1139	898	2037
		Gender 中的%	25.1%	16.5%	20.4%
合计		计数	4543	5457	10000
		Gender 中的%	100.0%	100.0%	100.0%

3.3. 相关性分析

绘制各个特征之间的相关系数表，红色越深则正相关性越显著，绿色越深则负相关性越显著。图 2 中显示，“流失”与“年龄”正相关关系最为密切，其次随着客户账户余额的增加，流失的可能性也会增大。“流失”与“购买本行产品数量”“信用度”均呈现出较为明显的负相关关系，这是对“购买本

行产品数量”这一特征而言，该数据集不平衡，购买产品数量在 1 种和 2 种的客户占了绝大多数，购买 3~4 种产品的客户所占比例非常小，因而对相关系数的影响很小。对“活跃用户”和“性别”特征，与决策变量“流失”的相关系数为负，即非活跃客户的流失率比活跃客户的流失率低，对于二分类特征，其特征值无大小关系。

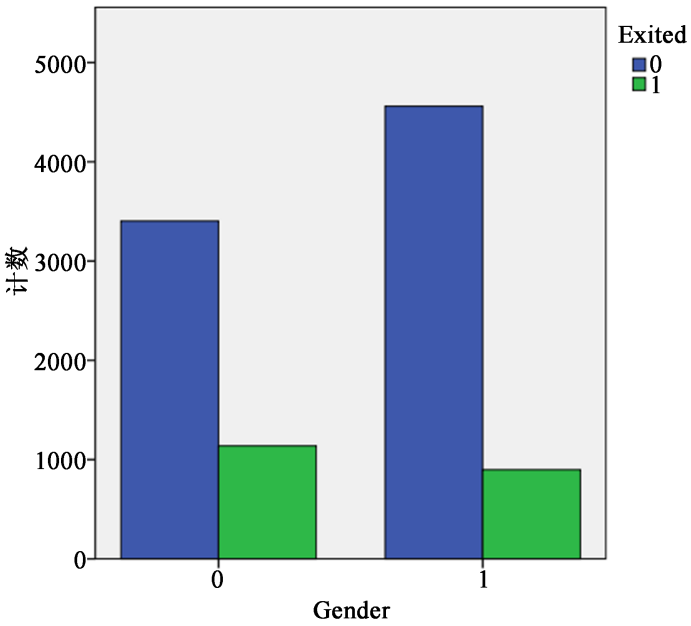

Figure 1. Comparison chart of the number of “gender” characteristics
图 1. “性别”特征数量对比图

	CreditScore	Gender	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary	Exited
CreditScore	1	-0.003	-0.004	0.001	0.006	0.012	-0.005	0.026	-0.001	-0.027
Gender	-0.003	1	-0.028	0.015	0.012	-0.022	0.006	0.023	-0.008	-0.107
Age	-0.004	-0.028	1	-0.01	0.028	-0.031	-0.012	0.085	-0.007	0.285
Tenure	0.001	0.015	-0.01	1	-0.012	0.013	0.023	-0.028	0.008	-0.014
Balance	0.006	0.012	0.028	-0.012	1	-0.304	-0.015	-0.01	0.013	0.119
NumOfProducts	0.012	-0.022	-0.031	0.013	-0.304	1	0.003	0.01	0.014	-0.048
HasCrCard	-0.005	0.006	-0.012	0.023	-0.015	0.003	1	-0.012	-0.01	-0.007
IsActiveMember	0.026	0.023	0.085	-0.028	-0.01	0.01	-0.012	1	-0.011	-0.156
EstimatedSalary	-0.001	-0.008	-0.007	0.008	0.013	0.014	-0.01	-0.011	1	0.012
Exited	-0.027	-0.107	0.285	-0.014	0.119	-0.048	-0.007	-0.156	0.012	1

Figure 2. Correlation coefficient heatmap
图 2. 相关系数热力图

3.4. 逻辑斯蒂回归模型

建模前对每个特征的所有数据使用 Z-Score 法进行标准化处理[5]，如公式(1)：

$$x^{*} = \frac{x - \mu}{\sigma} \tag{1}$$

其中 μ 为该特征所有样本数据的均值，σ 为该特征所有样本数据的标准差。经过处理的数据符合标准正态

分布，建模时可消除各类数据量纲不同的影响。对标准化后的数据进行逻辑斯谛回归，拟合结果如表 4 所示。

Table 4. Logistic regression model coefficient table
表 4. 逻辑斯谛回归模型系数表

	Estimate	Std. Error	z value	Pr (>|z|)	
(Intercept)	−1.63013	0.03047	−53.499	<2e−16	***
Score	−0.06303	0.02684	−2.349	0.0188	*
Gender	−0.27034	0.02687	−10.06	<2e−16	***
Age	0.76423	0.02676	28.561	<2e−16	***
Tenure	−0.04253	0.02683	−1.585	0.1129	
Balance	0.31479	0.02871	10.964	<2e−16	***
Num	−0.02104	0.02699	−0.78	0.4356	
Card	−0.01321	0.02678	−0.493	0.6217	
Act	−0.53943	0.0286	−18.86	<2e−16	***
Salary	0.02863	0.02702	1.06	0.2893	

*代表 p 值于 0.05，***代表 p 值小于 0.001。

由表 4 可知，有四个特征的 p 值远远小于显著水平 0.01，分别是“性别”“年龄”“账户余额”和“活跃客户”。系数值为正说明特征与决策变量呈正相关性，如“性别”的系数为−0.27034，“活跃客户”的系数为−0.53943，说明女性和不活跃客户流失的可能性更大。其次，“信用度”的显著性 p 值为 −0.06303 < 0.05，但是不小于 0.01，说明客户的信用度与其流失的显著性较为显著，系数值为负，说明信用度越低客户流失的可能性越大。

3.5. 模型修正

既然表 4 显示逻辑斯谛回归模型中有很多特征是不显著的，所以采用逐步逻辑斯谛回归的方式剔除某些不显著特征，逐步回归是一种选择“最优”预测模型的统计技术，它自动选择对响应变量影响最大的预测变量，并排除那些不显著的变量，用来识别对客户流失影响最大的因素组合，逐步回归的结果如表 5 所示。

Table 5. Logistic stepwise regression results
表 5. 逻辑斯谛逐步回归结果

	Df	Deviance	AIC	Estimate
<none>		8708.4	8.72E+03	−1.6296
Tenure	1	8710.9	8722.9	−0.0426
Score	1	8714	8726	−0.06328
Gender	1	8810.4	8822.4	−0.27
Balance	1	8846.8	8858.8	0.3211
Act	1	9091.6	9103.6	−0.53983
Age	1	9599.9	9611.9	0.76447

由表 5 可知，显著性特征为“年龄”“性别”“信用度”“账户余额”和“活跃客户”，同时，在简单逻辑斯谛回归模型的基础上，挖掘出了“任期”这一个显著性变量。比较可知，逐步回归是一边选择，一边剔除，确保每次引入新的变量之前回归方程中只包含显著性变量，直到既没有显著的解释变量选入回归方程，也没有不显著的解释变量从回归方程中剔除为止，最终可得到一个最优的变量集合。这样可以提高模型的可解释性和预测准确性，但需要大量的计算和迭代，并且具有很高的计算复杂度。

3.6. 模型评估

上述建模均是用所有样本数据进行模型拟合，得到影响客户流失的显著性特征，其模型混淆矩阵如表 6 所示。原始样本量共 10,000 条样本，准确预测未流失客户的样本共 7717 条，流失样本 353 条，预测为未流失实际上流失的样本 1638 条，得到未划分数据集时模型的预测准确率为 80.71%。

Table 6. The confusion matrix before partitioning the data set
表 6. 划分数据集之前的混淆矩阵

真实 \ 预测	0	1
0	7717	1683
1	246	354

为更好地评估模型性能，更好地检测模型的潜在漏洞，如过拟合等，将所有样本数据的前 5000 条样本划分为训练集，另外 5000 条样本作为测试集，计算得到测试集的混淆矩阵如表 7 所示，模型的准确率为 81.2%，相对于表 6 有所提高。原始数据量大，预测准确率却较低的原因多样，可能的原因例如：原样本数据集中包含数据噪声，导致在模型的表现下降，噪声可能是由于数据采集、处理或标记中的错误而引入的；原样本数据集在某些特征上存在数据不平衡的现象，当不同类别的样本数量不平衡时，模型可能会在某些类别上表现不佳，为解决这个问题，可尝试使用数据增强技术来增加少数类别的样本数量。

Table 7. The confusion matrix after partitioning the data set
表 7. 划分数据集之后的混淆矩阵

真实 \ 预测	0	1
0	3861	792
1	148	199

4. 案例结论与教学效果分析

4.1. 案例结论

在面对银行客户流失的问题时，理解哪些因素可能影响客户的行为是至关重要的。本文从三个层面进行分析：首先，通过探索性数据分析了解所有变量的分布情况以及与客户流失的相关关系，根据样本数据可知，客户的年龄、性别、信用度、是否为活跃客户以及账户余额都被认为是潜在的重要影响因素。具体分析来说，随着年龄的增长，人们的经济状况和社会地位可能更稳定，导致流失率降低；通常，男性和女性在金融行为上存在差异，这可能影响流失率；高信用度通常与良好的信用记录和较低的违约风

险相关，可能降低流失率；活跃客户更可能保持与银行的业务关系，从而降低流失率；较大的账户余额可能意味着客户对银行的信赖度较高，以及更好的财务状况，这可能影响流失率。其次，将“年龄”“性别”“账户余额”“活跃客户”等 9 个特征与因变量“流失”建立逻辑斯谛回归模型，定量地研究各个特征的显著性，并且对流失情况预测。分析发现除了上述较为显著的特征之外，客户在银行的任期也对流失率有显著影响，客户的银行任期可能反映其与银行的长期关系和满意度。一般来说，较长的任期可能意味着客户对银行有较高的信任和满意度，这可能导致较低的流失率。

4.2. 教学效果分析

在数据科学与分析这样的领域，案例学习是一种非常有效的学习方法，该案例结合实际应用背景，通过对逻辑斯蒂模型和逐步回归方法的运用，帮助同学们可以深入了解每个因素对银行客户流失的具体影响，让学生通过实践操作来掌握大数据分析的核心技术和方法，培养他们的实践能力和创新意识。同时，通过小组讨论和报告等形式，鼓励学生进行团队合作，并在现有分析的基础上进行创新，例如尝试使用不同的模型或方法，或从新的角度看待问题。同时案例教学过程中需要教师和学生共同参与，通过讨论、问答等方式进行沟通，也促进了教师与学生之间的互动，提高了教学效果和教育质量。

基金项目

校级教研项目(2023JGSZ25、2024JGSZ28)资助。

参考文献

[1] 庄荣文. 深入贯彻落实党的二十大精神以数字中国建设助力中国式现代化[EB/OL]. http://www.cac.gov.cn/2023-03/03/c_1679480312027134.htm?eqid=e731a8eb0022397300000002642e2c44, 2023-03-03.

[2] 全国 757 所高校成功申报数据科学与大数据技术专业教育部公布名单(2023 年) [EB/OL]. https://www.163.com/dy/article/I3BEIMMN0532N2UB.html, 2023-04-27.

[3] 高海燕. 基于数据挖掘的银行客户流失预测研究[D]: [硕士学位论文]. 西安: 西安理工大学, 2007.

[4] 一个极客. 银行客户流失[EB/OL]. https://www.datafountain.cn/datasets/4893, 2020-11-23.

[5] Shalabi, L., Shaaban, Z. and Kasasbeh, B. (2006) Data Mining: A Preprocessing Engine. *Journal of Computer Science*, **2**, 735-739. https://doi.org/10.3844/jcssp.2006.735.739