机器学习课程

大论文/机器学习实践报告

题目：基于机器学习的信贷风险预测

学号：2023110578

姓名：留祥鑫

1. 数据处理与特征工程

**数据来源与介绍：**

报告所使用的数据集来源于阿里云天池大赛-日常学习赛-【AI入门系列】金融守护者：金融风险预测学习赛。数据下载链接：

https://tianchi.aliyun.com/competition/entrance/531830/information

赛题以金融风控中的个人信贷为背景，要求根据贷款申请人的数据信息预测其是否有违约的可能，以此判断是否通过此项贷款。该数据来自某信贷平台的贷款记录，总数据量超过120w，包含47列变量信息，其中15列为匿名变量。为了保证比赛的公平性，将会从中抽取80万条作为训练集，20万条作为测试集A，20万条作为测试集B，同时对employmentTitle、purpose、postCode和title等信息进行了脱敏。由于测试集testA没有isDefault特征，本文对原训练集划分为新的训练集和验证集。

金融风控数据常见挑战是大量类别型特征和缺失值，因此必须进行特征工程，对复杂数据进行处理，从而降低过拟合风险，提高模型性能。

**数据特征说明：**

id 为贷款清单分配的唯一信用证标识

loanAmnt 贷款金额

term 贷款期限（year）

interestRate 贷款利率

installment 分期付款金额

grade 贷款等级

subGrade 贷款等级之子级

employmentTitle 就业职称

employmentLength 就业年限（年）

homeOwnership 借款人在登记时提供的房屋所有权状况

annualIncome 年收入

verificationStatus 验证状态

issueDate 贷款发放的月份

purpose 借款人在贷款申请时的贷款用途类别

postCode 借款人在贷款申请中提供的邮政编码的前3位数字

regionCode 地区编码

dti 债务收入比

delinquency\_2years 借款人过去2年信用档案中逾期30天以上的违约事件数

ficoRangeLow 借款人在贷款发放时的fico所属的下限范围

ficoRangeHigh 借款人在贷款发放时的fico所属的上限范围

openAcc 借款人信用档案中未结信用额度的数量

pubRec 贬损公共记录的数量

pubRecBankruptcies 公开记录清除的数量

revolBal 信贷周转余额合计

revolUtil 循环额度利用率，或借款人使用的相对于所有可用循环信贷的信贷金额

totalAcc 借款人信用档案中当前的信用额度总数

initialListStatus 贷款的初始列表状态

applicationType 表明贷款是个人申请还是与两个共同借款人的联合申请

earliesCreditLine 借款人最早报告的信用额度开立的月份

title 借款人提供的贷款名称

policyCode 公开可用的策略代码=1新产品不公开可用的策略代码=2

n系列匿名特征 匿名特征n0-n14，为一些贷款人行为计数特征的处理

**缺失值处理：**

|  |
| --- |
| 数据缺失值 |
| 图1.特征变量缺失情况 |

收集和整理数据的过程中数据缺失是难以避免的，但大量数据缺失会严重影响模型的性能，因此需要对缺失值进行处理。这里采用按照中位数填充数值型特征，按照众数填充类别型特征的方法进行填充。

**异常值处理：**

异常值也可能对模型性能产生较大影响。为提高信贷风险预测模型的准确性，需要进行异常值处理。本文用均方差检验异常值。在统计学中，如果一个数据分布近似正态，那么大约 99.7% 会在三个标准差范围内，超出该范围的数据视为异常值，并删除对应样本。

**分类变量编码：**

机器学习算法适用于处理数值型数据，但是对非数值型的文本或类别型数据适用性减弱，因此需要分类特征变量转换为数值型数据。本文首先对employmentLength和earliesCreditLine这两个时间特征进行转换；其次对grade和subGrade进行标签编码 (Label Encoder）；最后对homeOwnership, verificationStatus, purpose, regionCode使用独热编码（One-Hot Encoding）进行转换，并在转换后删除每个类别变量的第一列以避免共线性问题。

**归一化：**

对特征做归一化，可以让训练过程更好更快的收敛，同时避免某些特征对结果的影响过于显著，而其他特征的影响则被忽略的问题。

**特征扩展与选择：**

特征选择技术可以精简掉无用的特征，以降低最终模型的复杂性。

|  |  |  |  |
| --- | --- | --- | --- |
| 相关矩阵 |  |  |  |
| 图2.相关矩阵 |  |  |  |
| 与目标变量相关 |  |  |  |
| 图3.与目标变量的相关性可视化 |  |  |  |

经过相关性筛选与特征扩展，最终选择以下特征：

subGrade ：最强的预测器 (r=-0.262)

interestRate ：高相关性 (r=0.254)

term ：中期相关性 (r=0.175)

ficoRangeHigh ：关键信用评分指标

dti ：债务负担比率

verificationStatus ：收入验证状态

n3 ：保留的n系列中预测力最强的

revolUtil ：循环信用利用率

risk\_score=interestRate\*term/annualIncome ：综合风险指标

fico\_interest\_ratio=ficoRangeHigh/interestRate ：风险价格比

installment\_burden=installment/annualIncome /12 ：实际支付负担

subGrade\_employment= subGrade\*employmentLength+1 ：信用风险与就业交互

1. 技术调研

在信贷领域，与信贷风险预测相关的研究已经非常全面，各种数据挖掘和机器学习算法被广泛应用于风险预测和信用评估任务中。在金融风险预测领域（尤其是信贷风险预测），逻辑回归、支持向量机、GBDT、随机森林、深度神经网络等算法被广泛应用。其中，传统的模型如线性回归、逻辑回归和决策树等，由于其可解释性高，利于通过归因分析来进行业务决策，但它们的缺点也很明显，就是预测效果较差。而GBDT、随机森林和深度神经网络等新兴算法能够带来更高的预测性能，尽管他们各有优劣。

Lightgbm作为一种梯度提升决策树算法，具有高效、准确等特点，非常适合处理大规模数据和复杂特征的信贷风险预测。在广泛的表格数据分类/回归任务上，尤其是在中小规模到大规模数据集上，GBDT通常是精度最高的方法之一。它能够有效地挖掘特征之间的关联性，提取重要特征，快速训练模型。缺点在于高维稀疏数据的情况下它的表现可能不如DNN，而模型可解释性可能弱于随机森林。

随机森林的优势在于它的稳健性和可解释性。它通过Bagging算法降低方差，天然具有抗过拟合的特性；而且对异常值和缺失值不敏感，适用于金融风控数据。缺点是可能忽视弱特征组合效应，而且训练速度会慢于GBDT。

深度神经网络强大的特征表示和学习能力，能够自动从原始或预处理后的数据中学习高度非线性和复杂的特征交互与表示，且擅长处理高维稀疏数据；理论上，深度神经网络可以逼近任意复杂的非线性函数，能更好地拟合现实中极其复杂的信用风险决定因素之间的关系。但它具有可解释性差，对数据规模和质量要求高，训练和调优复杂的明显缺点。

在指标选择方面，本文选择AUC作为核心评估指标。信贷风控模型的核心目标不是要求对客户违约与否进行100%准确的绝对分类，而是尽可能将风险最高的客户准确地区分出来，并根据风险水平对客户进行有效排序，而AUC（ROC曲线下面积）衡量的是模型在所有可能的分类阈值下分类的能力，完美贴合了信贷风险预测模型的本质需求——对客户风险进行准确排序的能力，与风控场景的业务需求完全吻合。

1. 实验算法与实现

报告采用GBDT（用lightgbm实现）, Random Forest（用sklearn实现）, 以及深度神经网络（用Pytorch实现）三种算法，采用标准化评估流程，重点考察模型在AUC、准确率等关键风控指标上的表现，并通过训练时间和模型复杂度分析其实际应用价值，同时进行可视化。所有模型设置random\_state=42确保可复现性；GBDT通过样本权重参数，随机森林通过class\_weight='balanced'，DNN通过损失函数加权或过采样，来处理类别不平衡问题。Lightgbm基于梯度提升决策树框架，通过连续构建弱学习器（决策树）来修正前序模型的残差误差。实验以AUC作为优化目标，采用0.05的保守学习率保证稳定性，同时通过L1正则化、L2正则化确保模型泛化能力，并监控训练过程，早停防止过拟合；随机森林基于Bagging集成框架，通过构建多棵独立决策树并进行投票表决，实验控制树复杂度，采取有放回采样；深度神经网络则是端到端的深度特征学习架构，通过多层非线性变换提取高阶特征交互。通过全连接层自动学习特征表示，结合Dropout层和L2权重衰减，使用ReduceLROnPlateau调度器实现精细优化。

1. 实验结果与分析

为评估本章所建立信贷预测模型的性能，将GBDT、随机森林和DNN作为对比模型，具体实验结果和混淆矩阵分别如下所示。

|  |
| --- |
| 模型比较 |
| 图4.模型性能比较 |

|  |
| --- |
| GBDT混淆矩阵 |
| 图5.GBDT混淆矩阵 |

|  |
| --- |
| 随机森林混淆矩阵 |
| 图6.随机森林混淆矩阵 |

|  |
| --- |
| DNN混淆矩阵 |
| 图7.DNN混淆矩阵 |

由以上结果可知：Lightgbm模型在准确率和AUC两大指标上都显著优于随机森林；与深度神经网络相比虽然在AUC和准确率上相差不大，但从时间上来说Lightgbm远胜于深度神经网络算法。

1. 结论

本文首先针对金融风控数据大量类别型特征和缺失值的特点进行了数据处理与特征工程，并利用AUC指标对GBDT、随机森林和深度神经网络三种算法进行比较，发现GBDT在预测精度和效率上都有显著优势，并且具有良好的可解释性。利用GBDT（Lightgbm）模型，银行等金融企业可以更高效地进行贷款决策，降低时间和人力成本，有效降低违约风险，助力金融产业高质量发展。

六、参考文献

[1]庞素琳.违约风险下的信贷决策模型与机制[J].管理科学学报,2012,15(04):58-70.

[2]戴峥琪.基于可解释机器学习的个人信贷风险预测研究[D].吉首大学,2024.DOI:10.27750/d.cnki.gjsdx.2024.000040.

[3]周隽琪.深度学习可解释性在信贷违约风险预测模型中的应用研究[D].上海财经大学,2023.DOI:10.27296/d.cnki.gshcu.2023.002140.、

[4]庞烨.基于深度学习的信贷风险预测系统的实现[D].南京邮电大学,2022.DOI:10.27251/d.cnki.gnjdc.2022.000540.