基于 RoBERTa 和 LightGBM 的中文图书采选模型研究*

□钟建法　孟子正

摘要　在对智能图书采选模型构建方法进行综述和对相关机器学习算法进行介绍基础上，探索基于 RoBERTa 和 LightGBM 构建高校图书馆中文图书采选机器学习模型。分析模型的构建目标和研究框架，从数据来源与清洗、特征筛选与确定、衍生特征构建、基于 RoBERTa 模型的文本特征构造、数据编码等方面对特征工程进行详细描述，构建基于 LightGBM 的中文图书采选分类模型并进行模型评估，提出模型应用策略方案和后续研究建议，以期推进机器学习模型的应用发展和图书采选工作的智能化转型。

关键词　高校图书馆　图书采访　机器学习模型　RoBERTa　LightGBM
分类号　G253.1
DOI　10.16603/j.issn1002－1027.2025.01.010
引用本文格式　钟建法，孟子正.基于 RoBERTa 和 LightGBM 的中文图书采选模型研究[J].大学图书馆学报，2025，43(1)：82－92.

1　概述

智慧图书馆建设是中国式现代化新征程上图书馆事业高质量发展的重要方向，在建设智慧图书馆的新形势下，图书馆藏建设高质量发展的方向是走向数据化、智能化和精准化。图书采选是图书馆藏建设的核心工作，建立在对不确定需求预测的基础上，是对图书内容价值、读者文献需求和馆藏建设需要进行综合评估与权衡的复杂决策活动[1]。图书馆在推行读者决策采购模式以增进图书采选针对性的同时，探索基于大数据和人工智能技术的智能选书和模型应用是实现图书采选高效化和精准化的重要途径[2]。

由于中文图书年新增馆藏量大且以馆员选书为主，20 世纪 80 年代以来，中文图书采选模型研究与应用成为我国图书馆信息资源建设理论研究和实践关注的重点。在智能选书标准方面，20 世纪 90 年代，王积和[3]、张炎烈[4]在美国学者约翰·拉特里奇(John Rutledge)等提出的拉、斯氏智能选书标准[5]的基础上结合我国国情提出了改进应用方案，游丽华提出“图书采选参数表”[6]；21 世纪以来，国内研究着重从图书价值、学术水平、出版质量、读者需求、馆藏需要、书评引用等方面构建图书智能采选决策标准，同时大力发展欧美发达国家图书馆倡导的读者决策采购模式[7]。

在模型构建方面，学者们主要运用决策函数、层次分析法[8－9]和人工智能算法构建图书采选决策模型。层次分析法是中文图书采选模型构建较为常用的方法，王洁等选取学科类、学术水平等级、网络读者评价、出版社、作者、价格等 9 个评价指标[10]，钟建法等选取读者对象、图书类型、著作责任方式、学位职称、H 指数、学科出版社、出版馆藏比、学科专业设置等 17 个评价指标，构建层次分析法模型[11]。层次分析法模型基于专家调查和系统分析方法来处理复杂的图书采选决策问题，指标体系、权重设置和参数赋值主要依赖专家选择，对文本特征语义理解和处理能力较弱，存在客观性不足和预测准确性不高的缺陷[12]。

人工智能算法应用是构建图书采选决策模型的重点探索领域。傅立云等选取作者、出版社、有效借阅人数、平均借阅时长、是否符合馆藏结构等 11 个

* 福建省社会科学基金项目“基于机器学习的图书馆纸电图书协同采选模型构建及其应用研究”(项目批准号：FJ2023B111)的研究成果之一。
通讯作者：钟建法，ORCID：0000-0003-2197-0199，邮箱：jianfa@xmu.edu.cn。

指标，构建随机森林预测模型对图书馆配商的新书征订书目进行采购决策[13]；鞠静选取作者、出版社、出版时间、图书价格和图书类别等 5 个特征变量，构建基于决策树算法的高校图书馆图书采访决策模型[14]；周志强选择学科类别、出版社、读者需求度、出版时间等 8 个评估指标，构建基于遗传算法优化的 BP 神经网络和支持向量机图书采购模型[15]；王红等利用文本分词朴素贝叶斯分类方法提取馆藏图书的题名和出版社特征对图书流通趋势进行分类[16]，蔡丹丹基于遗传神经网络构建“热门书”和“冷门书”的采购推荐模型[17]。

现有研究在人工智能算法应用上进行了探索，但尚未形成能够深度融入实际图书采选业务，并具有较好预测效果的图书采选机器学习模型和辅助决策系统。近年来，随着数据科学与人工智能技术的发展，GPT、BERT 等自然语言处理模型和 LightGBM、XGBoost、CatBoost 等决策树集成学习模型得到广泛应用，结合这两类模型的自然语言处理优势和结构化数据处理优势是图书采选模型构建的重要探索方向。

基于以上分析，本文探索基于预训练语言模型 RoBERTa 和集成学习模型 LightGBM 构建高校图书馆中文图书采选机器学习模型，以 RoBERTa 作为文本特征提取器，以 LightGBM 作为最终分类预测模型，将 RoBERTa 提取的关键文本特征和其他反映图书内容质量与适藏性的特征，输入到分类性能较好的 LightGBM 模型中进行训练和预测，以期得到较好的模型预测效果。本文力求在以下三个方面有所突破和创新：

(1)针对中文图书征订书目规模大、字段多、文本内容丰富等特点，综合利用 RoBERTa 的文本理解能力与多特征信息挖掘能力和 LightGBM 的高效分类性能与预测能力，提升中文图书采选分类预测的效率和准确性。

(2)从实用性出发构造实验样本和设计模型用途，增强模型与实际图书采选业务的契合度。实验样本采用作为高校图书采访主要书目来源的馆配商征订书目，通过分析模型预测的图书入选馆藏概率，为图书采选提供荐藏、选藏和不藏三种模型应用策略方案。

(3)在文本数据处理上，不仅利用 RoBERTa 输出图书的读者对象、作者和内容等相关文本语义特征的预测概率，而且结合图书馆员的专业知识，利用传统关键词匹配方法构造主要读者对象、图书类型等衍生特征，促进文本特征分类准确性与可解释性的共同提升。

2 相关机器学习算法简介

2.1 RoBERTa 简介与运用

BERT(Bidirectional Encoder Representation from Transformers)是谷歌公司于 2018 年开发的预训练语言模型，它通过 Transformer 架构，使用自监督学习方式，在大规模语料库中通过掩码语言模型(Masked Language Model，MLM)和下句预测(Next Sentence Prediction，NSP)两个任务进行预训练和学习文本的语义表示，在文本分类等多项自然语言处理任务中有很好的效果[18]。RoBERTa(Robustly optimized BERT pretraining approach)是 BERT 的优化和改进版本，它使用更大的参数规模和训练数据，利用动态掩码替代静态掩码，并取消了下句预测任务，将 BERT 预训练模型的性能提升到一个新的高度[19]。

基于 RoBERTa 的下游自然语言处理任务，通常采用微调(Fine-tuning)的方式来实现。对于图书文本分类任务，在 RoBERTa 模型基础上添加一个线性层和 Softmax 层，将 RoBERTa 解码器输出的向量转换为样本类别的概率分布，通过交叉熵损失函数计算分类损失。然后，通过反向传播算法计算梯度，运用梯度下降法更新模型参数，迭代优化模型性能，提升模型对文本分类的准确性。

RoBERTa 通过学习文本语义的上下文表示，能较好地提取关键语义信息。在模型应用中，从文本特征对图书采选决策的重要性和防止因拼接文本过长导致模型训练时显存溢出两方面考虑，将图书文本数据进行区分并拼接为读者相关文本、作者相关文本和内容相关文本，然后将三类文本输入到相应的 RoBERTa 模型中进行微调和预测，并将 RoBERTa 输出的文本特征分类预测概率值，输入到 LightGBM 模型中完成后续的图书采选分类任务。

2.2 LightGBM 算法简介与运用

LightGBM(Light Gradient Boosting Machine)是基于梯度提升决策树(GBDT)的集成学习算法模型[20]，它采用直方图算法寻找最优分割点，采用单边梯度抽样算法(GOSS)减少数据量，采用互斥特征捆绑算法(EFB)降低特征维度，采用带有深度限制

的按叶子生长(Leaf-wise)策略得到更好的精度。由于 LightGBM 运行速度快，预测效果较好，擅长处理结构化数据和数值型特征，且能较好地处理高维离散特征和缺失值，是机器学习领域中非常受欢迎的一个集成学习模型[21]。

在融合模型中，以 RoBERTa 作为文本数据的特征提取器，结合书目信息特征以及反映馆藏发展需要和适藏性的衍生特征，构建一张规范的结构化数据表。通过 LightGBM 模型对结构化数据进行学习和训练，最终形成一个既能有效利用书目文本数据和结构化数据，又能符合图书采选业务逻辑的中文图书采选模型，促进智能图书采选应用的发展。

3 模型构建目标和研究框架

3.1 模型构建目标

模型构建目标是为高校图书馆针对内地中文图书征订书目开展图书采选工作，提供较为可靠和有效的分类预测机器学习模型支持和实际采选应用方案。图书馆配商以及部分出版社提供的中文图书征订书目，是高校图书馆采购纸质中文图书的主要书目来源和重要渠道。模型基于主流业务需求，选取一定时期的、经过标注的大规模征订书目作为实验数据集，选择和构造反映图书内在属性和外在属性的特征变量进行建模和训练，充分学习实验数据与样本标签之间、图书价值与适藏性和采选偏好之间的有机联系，构建具有较好预测性能的图书采选二分类模型并进行可靠性和泛化能力评估。同时，对模型预测的征订书目入选馆藏概率进行区间分析，提供荐藏、选藏和不藏三种采选应用策略方案。

3.2 模型研究框架

模型研究框架和技术路线如图 1 所示：首先，按照模型构建要求采集原始数据，并对原始数据进行预处理和匹配合并；其次，依据一定原则对特征进行初步筛选，排除无效和不予应用的特征，然后对拟用特征进行数据清洗和规范化；再次，对于文本数据利用 RoBERTa 提取文本特征，对于类别型特征和衍生特征使用标签编码方法和专家评分法进行数据编码，数值型特征数据则直接利用，最终形成一张规范的结构化数据表；最后，构建 LightGBM 分类模型进行训练和预测，利用测试集数据对模型进行评估和结果分析，提出模型应用策略方案和后续研究建议。

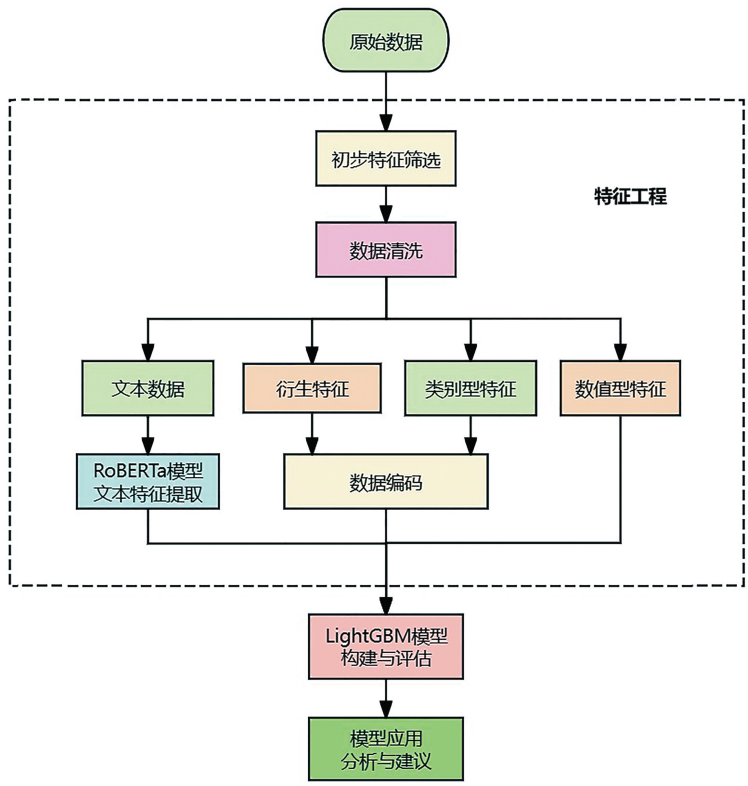

图 1 模型研究框架和技术路线

4 特征工程

4.1 数据集来源与划分

本研究以 2017－2022 年内地图书出版单位出版的图书为范围，在中文图书馆配商湖北三新文化传媒有限公司和数据服务提供商杭州麦达电子有限公司的支持下，采集到数据规模大且相对完整的内地中文图书征订书目，在删除没有 ISBN 的图书和书目去重后，保留 892366 种中文图书征订书目作为数据集。馆藏分析数据选取厦门大学图书馆（以下简称厦大馆）集成管理系统中导出的馆藏图书数据并与中文图书征订书目进行匹配，匹配后的厦大馆藏图书数据为 248547 种。数据标注如表 1 所示，将征订书目中落选厦大馆藏样本标记为 0，入选厦大馆藏样本标记为 1，其中落选样本和入选样本占征订书目总种数的比例分别为 72％和 28％，样本不平衡现象明显。

对于数据集中的 892366 条样本数据，按照 8∶1∶1 的比例划分为训练集、验证集和测试集，训练集样本量为 713892 条，验证集和测试集样本量均为 89237 条。训练集用于模型训练，验证集用于参数调节、特征选取和模型择优，测试集用于模型评估和结果分析。

表 1 数据集标签分布和描述

类别标签	类别名称	类别描述	征订书目种数	种数占比
0	落选	征订书目中落选厦大馆藏样本	643819	72％
1	入选	征订书目中入选厦大馆藏样本	248547	28％

4.2 特征筛选与确定

特征工程是使用领域专业知识选择、提取、操作和转换数据的过程，目的在于生成合适的特征用于模型训练和预测。模型特征筛选经历两个阶段：一是初步筛选阶段。以文献归纳和实践分析为基础，以适用性、完整性、差异性和可解释性为特征筛选原则，全面梳理和分析影响图书采选决策的图书特征、作者水平、出版质量、评价信息、馆藏建设需要和读者利用需求等因素，形成初步特征筛选方案；二是最终确定阶段。以模型性能优化为特征筛选原则，在模型训练和测试中对特征进行调整和确定。模型最终确定使用的特征为 17 个，特征类型、特征数量和特征名称见表 2。

表 2 模型最终确定使用的特征

特征类型	特征数量	特征名称
书目信息特征	10	价格、页码、开本、分类号、出版社、著作方式、语种、学科主题 a、学科主题 j、学科主题 x
衍生特征	4	学科出版社分区、馆藏收藏级别、主要读者对象、图书类型
RoRERTa 文本特征	3	读者文本 RoBERTa 预测概率、作者文本 RoBERTa 预测概率、内容文本 RoBERTa 预测概率

备注：学科主题 a、学科主题 j 和学科主题 x 分别对应中国机读目录格式（CNMARC）的 606a、606j 和 606x 子字段。

4.3 数据清洗

初步筛选特征后，需要对原始数据进行清洗，包括删除非内地图书出版单位出版的图书书目，更正 ISBN 号著录错误，对分类号和出版社进行统一和归并，以及对价格、页码、开本等进行规范。

（1）价格、页码和开本数据清洗。这三类数据在原始数据集中以文本数据方式存储，需要转换为数值型特征。通过结合正则表达式，重复匹配与标准模式不符合的样本并对其进行修正，得到价格、页码和开本的数值型数据，对无法提取数值的样本则设置为空值。

（2）分类号归并。采访馆员选书判断和智能采选模型应用通常以《中国图书馆分类法》（以下简称中图法）细分类目为分析基础。根据高校图书馆图书采选工作实际，选取中图法大类、二级类目和部分符合业务需求的三级类目作为模型使用的细分类目，这些细分类目对应的分类号共计 220 个。

（3）出版社名称规范。原始数据中的出版社名称，存在全称和简写并存、简体和繁体并存、著录书写有误、不同时期更名等问题，需要进行统一和规范。本研究整理了同一出版社在不同时期和更名前后使用的出版单位前缀，形成“图书类出版单位前缀与名称规范对应表”，按出版单位前缀和统一的出版社名称对出版社进行规范。

4.4 基于图书采选业务需要的衍生特征构建

衍生特征是在分析图书采选业务逻辑和业务数据的基础上，依据图书采选业务需求和模型决策需要而构建的新特征。衍生特征包括反映图书出版质量的学科出版社分区、反映馆藏发展需要的馆藏收藏级别、反映读者需求的主要读者对象和图书类型等 4 个特征，全部为人工提取的类别型特征。构建这些衍生特征，目的在于更好地体现图书采选业务特性，提升模型表现和更好地解释模型。

(1)学科出版社分区特征构建。学科出版社分区包括 1 区(核心区)、2 区(次要区)、3 区(相关区)、4 区(无关区)等 4 个类别。学科出版社分区表的构建，以 2017 至 2022 年为图书出版时间范围，以 220 个中图法类目(包括 22 个一级类目和 198 个细分类目)为分析对象，对厦大馆藏图书种数和借阅图书种数，综合运用熵值法、布拉德福区域分析法和遍历法，计算并形成由 220 个中图法类目和 579 个出版社组成的出版社分区对应表，并按相应类目与出版社，将对应表结果匹配到征订书目的学科出版社分区特征中。

(2)馆藏收藏级别特征构建。馆藏收藏级别包括特藏级、研究级、学习级、基础级、最低级和不收藏等 6 个类别，通过构造馆藏收藏级别映射表进行特征赋值。厦大馆藏收藏级别映射表由 2102 个中图法类目及相应馆藏收藏级别构成，映射表中的类目和收藏级别，由采访馆员在充分调研学校学科专业设置和建设方案基础上，依据馆藏发展政策的规定商议和确定。

(3)主要读者对象特征构建。主要读者对象包括 164 个类别，类别设置以《中华人民共和国职业分类大典(2022 年版)》(公示稿)的职业小类为主要依据，同时结合征订书目使用对象附注字段的内容描述和采选业务需求而定。构建主要读者对象特征，目的在于为样本分类确定一个最主要的使用对象，方法是预先构造主要读者对象类别与关键词对应表，然后在征订书目的使用对象附注以及一般性附注、丛编项、学科主题 x 等字段中查找和匹配相应的关键词进行特征赋值。

(4)图书类型特征构建。图书类型包括学术专著、重点图书、研究报告、重点教材、一般教材、教参、书集、史料、地方志、会议录、论文集、年鉴、词典、手册、指南、培训考试用书等 29 个类别，通过预先构造图书类型的类别与关键词对应表，然后在征订书目中的书名、一般性附注、丛编项、学科主题等字段查找和匹配相应关键词进行特征赋值。

4.5 基于 RoBERTa 模型的文本特征构造

4.5.1 实验环境

文本特征构造使用阿里巴巴集团通义实验室在魔搭社区上发布的“RoBERTa 预训练模型－中文－base”模型，结合 transformers 库进行微调。RoBERTa 微调往往需要借助 GPU 进行加速训练，在实验环境上，使用 OneThingAI 算力平台的租用服务器，该服务器的 GPU 类型为 RTX 4090，GPU 数量为 1，GPU 显存为 24GB，CPU 核数为 16 核，内存容量为 64GB。

4.5.2 文本数据构造

依据构造需要，三类 RoRERTa 文本特征的文本数据分别由样本书目中不同字段的文本信息拼接而成。其中，内容相关文本由书名、丛编项、一般性附注、学科主题 a、学科主题 j、学科主题 x、提要文摘等书目字段的文本拼接而成；作者相关文本由责任者、责任者附注、分类号、学科主题 a、学科主题 j、学科主题 x 等书目字段的文本拼接而成；读者相关文本由使用对象附注、学科主题 a、学科主题 j、学科主题 x 等书目字段的文本拼接而成。三类拼接文本经过转换与预处理后，分别输入到相应的 RoBERTa 模型中进行微调和预测，即可输出介于 0 和 1 之间的相应文本的入选馆藏预测概率值。

4.5.3 文本转换与预处理

原始文本数据在输入 RoBERTa 模型之前，需要使用分词器(Tokenizer)转换为 RoBERTa 可以处理的格式和词元(token)序列。Tokenizer 将每类文本编码为三个张量：input_ids、token_type_ids 和 attention_mask，这些张量之后将输入到 RoBERTa 模型中进行计算。输入序列最大长度(max_length)是 Tokenizer 的主要参数，Tokenizer 将 token 数量大于 max_length 的句子进行截断，小于 max_length 的句子进行填充。通过观察图 2 文本长度直方图中的三类文本数据的长度分布，将读者、作者和内容三类相关文本的 max_length 分别设置为 128、256 和 512 个 token，以便控制模型计算量，确保高效处理文本数据。

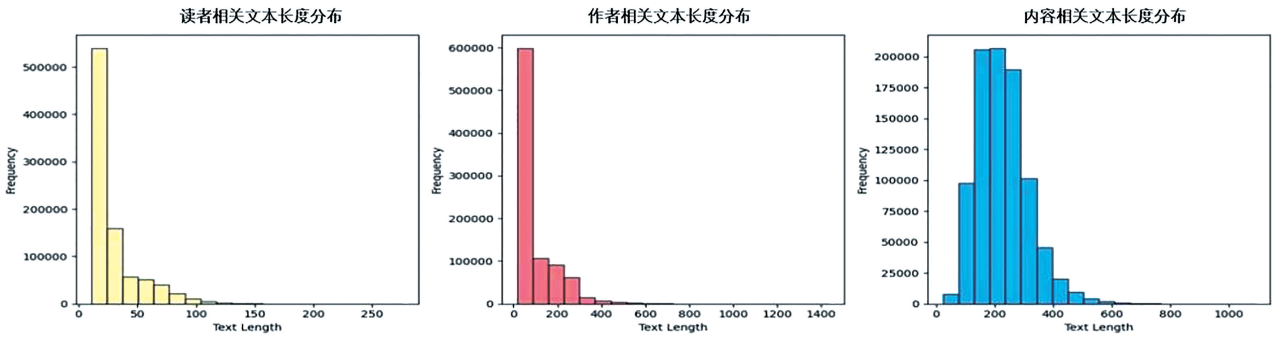

图 2　读者、作者和内容三类相关文本的文本长度直方图

4.5.4　RoBERTa 微调和预测

首先，在 RoBERTa 模型的基础上融入一个由线性层和 Softmax 层组成的分类层，线性层将 RoBERTa 输出的高维特征有效映射至标签空间，以实现分类预测；Softmax 激活函数将线性层输出的两个样本类别的分数转换为概率（概率之和为 1），应用于处理输入文本的二分类预测任务。

其次，调用 Trainer 函数对 RoBERTa 模型进行微调。Trainer 函数是 transformers 库中的一个高级 API，用于微调 RoBERTa 模型以适应特定的文本分类任务，并能起到简化模型训练过程的作用。Trainer API 根据配置的超参数迭代训练数据集，不断调整 RoBERTa 模型的参数，提高模型在给定任务上的性能。Trainer 训练过程的主要超参数配置见表 3。

表 3　读者、作者和内容三个 RoBERTa 模型的训练过程主要超参数

Model	optimizer	batch_size	num_train_epochs	learning_rate	weight_decay
读者	AdamW	64	5	0.00002	0.01
作者	AdamW	64	5	0.00002	0.01
内容	AdamW	32	5	0.00002	0.01

将训练集和验证集中经过转换和预处理后的读者、作者和内容三类文本数据以及对应分类标签，输入到 RoBERTa 模型中通过 Trainer API 执行监督训练任务。从图 3 可以看出，三个 RoBERTa 文本分类模型的训练集和验证集的损失变化，随着训练轮次（Epoch）的增加，训练集的损失（Training Loss）一直减小（见图 3 中从左上角延伸到右下角的折线），而验证集的损失（Validation Loss）先减小后增大，三个模型均在第三轮训练时达到验证集的最好效果。由于目标任务是借助 RoBERTa 将文本转化为入选馆藏的预测概率并作为后续模型的输入特征，因此，要求训练集与验证集的分布尽可能一致，在图 3 中，第 2 轮训练的验证集损失较小且训练集与验证集的损失最为接近，所以选择第 2 轮的训练结果作为最终结果。

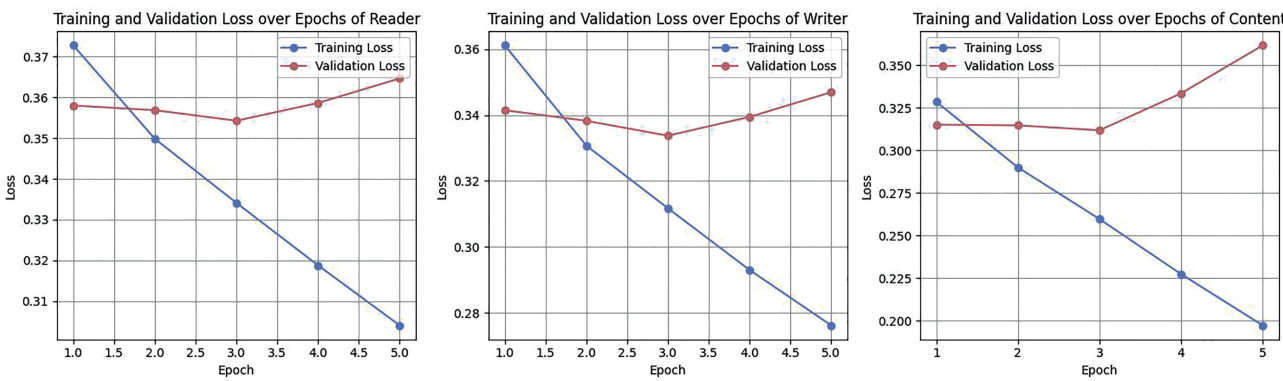

图 3　读者、作者和内容三个 RoBERTa 文本分类模型在微调过程中的训练集和验证集损失变化情况

最后，对读者、作者和内容三类相关文本，分别使用对应的微调后的 RoBERTa 模型进行二分类预测，模型最终输出的两个样本类别的分数经由 Softmax 函数转换为概率，并选取样本类别为“入选”的概率作为提取到的文本特征，完成读者文本 RoBERTa 预测概率、作者文本 RoBERTa 预测概率和内容文本 RoBERTa 预测概率三类文本特征的构建。

4.6 数据编码

数据集的类别型特征和衍生特征，采用标签编码(Label Encoding)和专家评分法进行数据编码。标签编码将类别编码为唯一的整数，是机器学习中一种常见的编码方式。对于类别较多的离散值，首先使用标签编码，将训练集中的每个类别编码为一个唯一的整数；其次，对于验证集和测试集，若其类别出现在训练集中，则编码为训练集中该类别对应的整数，若其类别未出现在训练集中，则编码为空值。对于类别较少的离散值和衍生特征，采用专家评分法编码，构造专家评分映射表对离散特征进行赋值。

经过数据清洗、特征衍生和离散特征编码、RoBERTa 文本特征提取，最终将全部样本数据编码为数值类型，形成一张规范的结构化数据表。在表 4 中，以验证集样本数据作为示例，对编码前和编码后的数据进行对照和描述。

表 4 验证集数据编码示例

特征名称	非空值总数	编码前数据	编码后数据
价格	89235	CNY55.00	55
页码	85290	355 页	355
开本	85886	26cm	26
语种	74381	chi	11
出版社	89237	北京大学出版社	564
分类号	89237	TP	206
著作方式	83369	主编	9
学科主题 a	85215	Python	68
学科主题 j	35873	教材	161
学科主题 x	62272	程序设计	11408
主要读者对象(衍生特征)	89237	高等学校师生	8
图书类型(衍生特征)	60563	重点教材	9
学科出版社分区(衍生特征)	89237	2 区(次要区)	6
馆藏收藏级别(衍生特征)	89237	研究级	9
读者文本 RoBERTa 预测概率	89237	—	0.71282780
作者文本 RoBERTa 预测概率	89237	—	0.45461926
内容文本 RoBERTa 预测概率	89237	—	0.92195960

5 基于 LightGBM 的分类模型构建与评估

5.1 LightGBM 二分类模型构建

模型构建目的是为高校图书馆开展内地中文图书征订书目的图书采选工作，提供较为可靠和有效的分类预测结果。从采选业务需求来说，期望构建的模型是一个整体预测效果较好且更加关注入选预测准确率的、与实际图书采选业务相契合的二分类预测模型。

由于数据集类别标签为 0 和 1 的样本比例分别为 72% 和 28%，正负样本不平衡现象明显，在 LightGBM 模型中设置数据集的类别权重值时，需要将 0 和 1 样本损失权重比例设置为 0.28 和 0.72

（与样本类别比例成反比）以缓解类别不平衡问题。在 LightGBM 模型中，输入编码后的结构化数据进行模型训练和预测，并取预测为入选的概率值作为输出结果。在模型训练中，使用早停的方式，当验证集表现在连续 50 轮不提高时即停止训练，以防止模型过拟合，最终模型在第 281 轮时停止训练，此时训练集的 multi_error 为 0.0904242，验证集的 multi_error 为 0.130000。经过多次调参尝试的 LightGBM 模型的其他超参数设置见表 5。

表 5　LightGBM 模型超参数设置

参数名称	参数设置	参数名称	参数设置
objective	multiclass	num_leaves	255
num_class	2	min_data_in_leaf	1024
metric	multi_error	learning_rate	0.1
alpha	1	subsample	0.8
boosting_type	gbdt	colsample_bytree	0.95

5.2　模型性能评估

5.2.1　特征重要度分析

特征重要度有助于理解 LightGBM 模型在决策过程中，哪些特征起到关键作用。在图 4 的特征重要度排序中，出版社和页码最高，三类 RoBERTa 文本特征、学科主题、价格、分类号次之，衍生特征和著作方式再次之，语种最低。总体而言，特征重要度排序较好地反映了馆员选书判断的决策逻辑，能够体现出版社、页码、读者相关文本、作者相关文本、内容相关文本和学科主题等特征对于图书采选判断的重要性。

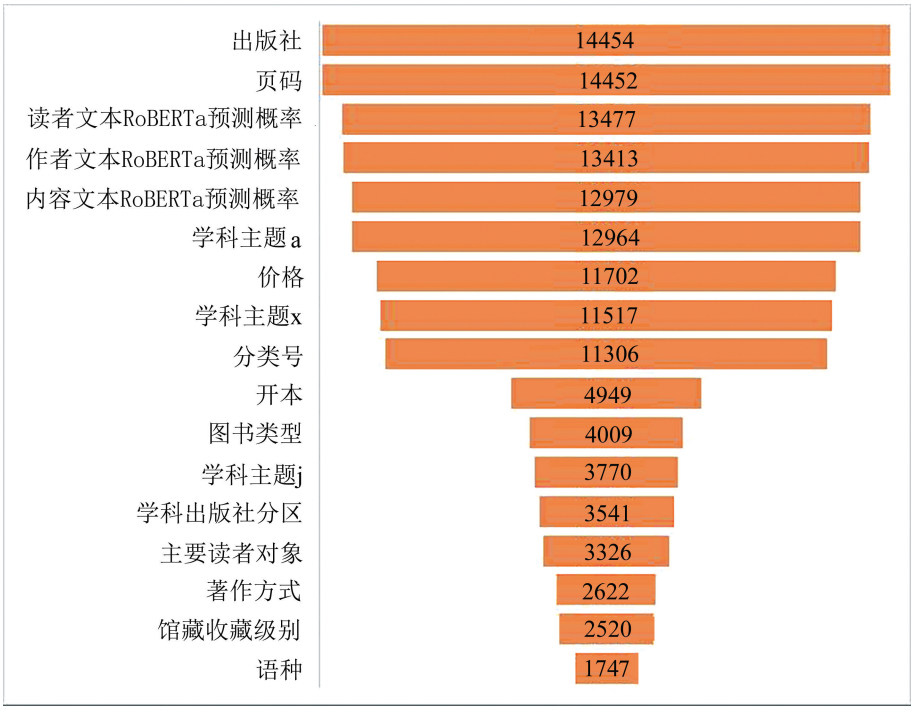

图 4　LightGBM 模型特征重要程度

5.2.2　ROC-AUC 和混淆矩阵评估

对于类别不平衡的二分类问题，综合运用 ROC－AUC 值、准确率和召回率等评价指标来评估模型的性能和效果较为合适。ROC－AUC 值通过遍历阈值的方式，以 ROC 曲线下的面积作为得分，更适合样本不平衡的二分类问题评估；准确率为预测正确的结果占总样本的比率，是衡量整体预测准确程度的指标；召回率又称查全率，是指实际入选馆藏样本中被正确预测为入选的样本比例，召回率越高，代表实际入选馆藏样本被预测出来的概率越高。

LightGBM 模型的性能评估以测试集结果为基准，测试集的 ROC－AUC 值为 0.942（ROC 曲线见图 5），准确率为 0.862，召回率为 0.884，整体表现较好。ROC－AUC 值越大，分类模型性能越好，按照 AUC 的性能度量一般标准，当 ROC－AUC 值介于 0.85 至 0.95 之间时，模型效果好。总体来看，模型的 ROC－AUC 值、准确率和召回率均较高，预示着模型有较好的预测性能和泛化能力，保证了图书采选的分类效果。

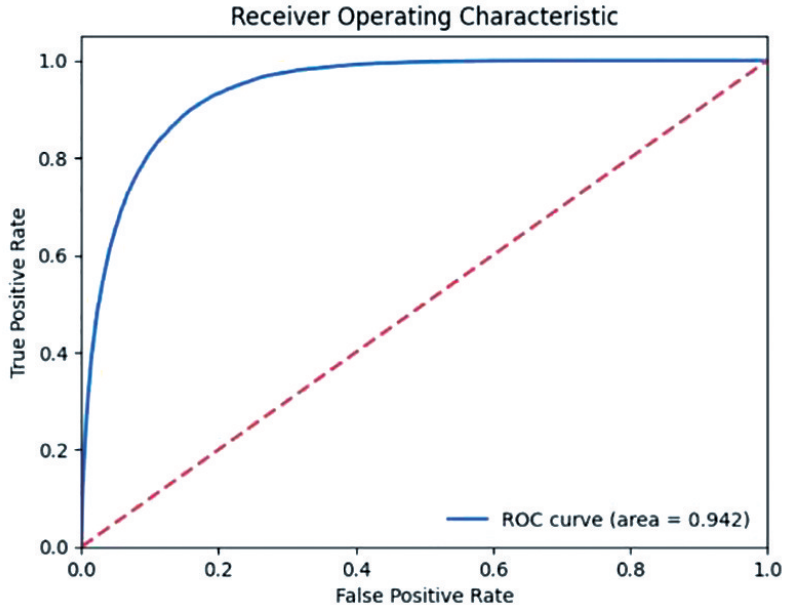

图 5　LightGBM 模型测试集 ROC 曲线

混淆矩阵是机器学习中总结分类模型预测结果的分析表，以清晰的视觉方式展示模型的预测结果与真实标签之间的关系，可帮助评估模型的分类性能，了解模型的强项和弱点。通过表 6 的混淆矩阵可以分析得出，测试集中实际入选馆藏样本共计 24831 种，其中被正确预测为入选的样本为 21948 种，召回率为 0.884，基本达到图书采选决策对模型预测性能的要求。同时，模型还在 64406 种实际落选馆藏样本中，挑选出 9414 种样本预测为入选，供采访馆员选择判断。

表 6　LightGBM 模型测试集混淆矩阵

实际类别	预测类别	
	预测入选	预测落选
实际入选	21948	2883
实际落选	9414	54992

6　模型应用策略方案和后续研究建议

6.1　模型应用策略方案

模型的应用价值体现在模型的预测结果能够作为实际图书采选工作的决策依据，通过分析模型的入选概率预测值能够为图书采选提供有效的应用策略和方案。从表 7 可以看出，在测试集入选概率预测值的高分值区间（预测值介于 0.9 到 1 之间），区间实际入选馆藏样本数量为 13132 种，占区间所有样本数量的 88%，占测试集所有实际入选馆藏样本 24831 种的 52.9%，入选预测的精准率和规模数量都较高，可将该区间样本的采选策略方案命名为“荐藏”；在入选概率预测值的中分值区间（预测值不低于 0.5 且不高于 0.9），入选预测精准率和效果适中，该区间的样本需要慎重采选，采选策略方案可命名为“选藏”；在入选概率预测值的低分值区间（预测值低于 0.5），实际入选馆藏样本仅为 2883 种（内含部分不适合入藏样本），而实际落选馆藏样本有 54992 种（占区间所有样本数量 57875 种的 95%），可见，该区间实际落选馆藏样本的预测精准率和规模数量极高，采选策略方案可命名为“不藏”。通过以上分析，依据模型的入选馆藏概率预测值进行区间划分，可以为实际图书采选工作提供荐藏、选藏和不藏三种模型应用策略方案。

表 7　测试集入选概率预测值区间的实际入选馆藏样本数量及比率

入选概率预测值区间	区间实际入选馆藏样本数量	区间所有样本数量	区间实际入选馆藏样本占区间所有样本的比率	区间实际入选馆藏样本占所有实际入选馆藏样本的比率	采选应用策略方案
[0.9－1]	13132	14986	88%	52.9%	荐藏
[0.5－0.9)	8816	16376	54%	35.5%	选藏
[0－0.5)	2883	57875	5%	11.6%	不藏
合计	24831	89237	28%	100%	

备注：区间所有样本数量为入选概率预测值区间的实际入选馆藏样本和实际落选馆藏样本的数量总和。

6.2　模型后续研究建议

（1）目前衍生特征主要通过构建关键词匹配表方式来赋值，工作量较大，复杂度较高，后续建议采用文本聚类与人工结合形式，将关键词匹配表转化为关键词向量表，以提高泛化能力。

（2）由于数据获取困难或提取复杂等原因，对图书采选决策具有一定影响的其他特征，如 H 指数、图书被引量、网络评分等，尚未应用到模型中。增加这些新特征可能会进一步提升模型的预测效果，后续建议通过爬虫技术及命名实体识别等技术提取新特征。

（3）由于征订书目中包含大量的文本信息和文本特征，大模型具有通用知识且具有更强的语义理解能力，后续可尝试运用大模型微调方式构建性能更佳的分类预测模型。

7　结语

本研究在借鉴和比较现有研究成果基础上，探索基于预训练模型 RoBERTa 和集成学习模型 LightGBM，构建高校图书馆中文图书采选分类预测模型，应用于高校图书馆的征订书目采选工作。通过借助 RoBERTa 的文本理解能力和 LightGBM 的高效分类性能，较好地解决了现有智能图书采选模型在图书文本特征提取和高维离散特征与缺失值处理上遇到的困难，最终形成一个既能有效利用书目文本数据和结构化数据，又能深度融入实际图书采访业务的、具有较好预测效果的中文图书采选模型，并提供荐藏、选藏和不藏三种有效的模型应用策略方案为实际图书采选工作服务，从而推进图书采选机器学习模型的研究和实践进展。由于课题组资源有限和外部数据获取不易，特征构造和模型构建仍有许多不足，期待后续研究者进一步完善和深化，共同推动机器学习模型的应用发展和图书采选工作的智能化转型。

参考文献

1　王红，雷菊霞. 人工智能图书采访模式设计及流程运维[J]. 图书馆学研究，2018(5)：71－76.
2　蔡迎春. 智能选书：图书馆精准采购实现策略[J]. 数字图书馆论坛，2021(6)：50－55.
3　王积和. 拉、斯氏选书标准及其改进方案[J]. 大学图书馆学报，1993(2)：4－6.
4　张炎烈. 国内外智能选书研究进展[J]. 图书情报工作，1994(2)：13－17.
5　Rutledge J，Swindler L. The selection decision：defining criteria and establishing priorities[J]. College and Research Libraries，1987(48)：123－131.
6　游丽华. 高校图书馆中文图书采访模型的初步研究[J]. 图书情报工作，1995(4)：33－36.
7　Tyler D C. Patron-driven purchase on demand programs for printed books and similar materials：a chronological review and summary of findings[J]. Library Philosophy and Practice，2011(6)：108－127.
8　资芸，钟叔玉，董毅明. 高校图书馆图书采访决策模型研究[J]. 情报杂志，2007(6)：145－147.
9　蔡迎春. 基于层次分析法的学科图书采购模型构建及实证分析[J]. 图书情报工作，2010，54(21)：36－39.
10　王洁，黄晓琴. 基于层次分析与边际效益计算的高校图书馆中文图书荐购模型[J]. 情报理论与实践，2016，39(4)：108－113.
11　钟建法，陈娟，李灿元，等. 高校图书馆图书采访决策模型研究[J]. 大学图书馆学报，2021，39(5)：38－47.
12　卞丽琴，陈峰. 基于人工智能的图书订购策略分析[J]. 图书馆杂志，2015，34(8)：39－43.
13　傅立云，胡芸. 基于随机森林的图书采购决策模型[J]. 情报探索，2020(5)：34－39.
14　鞠静. 基于决策树算法的高校图书馆图书采访决策模型研究[D]. 保定：河北大学，2021.
15　周志强. 基于混合智能算法的高校图书馆图书采购模型研究[D]. 呼和浩特：内蒙古大学，2019.
16　王红，王雅琴，黄建国. 基于文本分词朴素贝叶斯分类的图书采访机制探索[J]. 现代情报，2021，41(9)：74－83.
17　蔡丹丹. 基于遗传神经网络的图书采购推荐模型研究[J]. 图书

馆研究与工作，2022(5)：38－42.
18 赵旸，张智雄，刘欢，等. 基于 BERT 模型的中文医学文献分类研究[J]. 数据分析与知识发现，2020，4(8)：41－49.
19 梅侠峰，吴晓鸽，黄泽民，等. 融合 RoBERTa 的多尺度语义协同专利文本分类模型[J]. 计算机工程与科学，2023，45(5)：903－910.
20 Ke G L，Meng Q，Finley T，et al. LightGBM：a highly efficient gradient boosting decision tree[C]//NIPS'17：proceedings of the 31st international conference on neural information processing systems. New York：Curran Associates Inc.，2017：3149－3157.
21 邢红梅，陈欣，王慧. 基于 LightGBM 模型的文本分类研究[J]. 内蒙古工业大学学报(自然科学版)，2020，39(1)：52－59.

作者贡献说明：
钟建法：选题与框架设计，数据获取、清洗、分析与模型构建，论文调研、撰写、修改与审定
孟子正：数据清洗、模型构建与实验，论文撰写与修改

作者单位：钟建法，厦门大学图书馆，福建厦门，361005
孟子正，厦门大学经济学院，福建厦门，361005

收稿日期：2024 年 5 月 8 日
修回日期：2024 年 5 月 30 日
（责任编辑：李晓东）

Research on Chinese Book Acquisition Model Based on RoBERTa and LightGBM

ZHONG Jianfa　MENG Zizheng

Abstract: Exploring intelligent book selection and model application based on big data and artificial intelligence technology is an important way of high-quality development of library collection construction. Based on the review of the intelligent book selection standards and the construction methods of book selection model, this paper introduced the functions and roles of RoBERTa model and LightGBM algorithm, and explored the construction of a machine learning model for Chinese book acquisition in university libraries based on RoBERTa and LightGBM. The purpose is to provide a reliable and effective classification prediction model and practical selection application scheme for university libraries to carry out book acquisition based on Chinese book subscription bibliography, promoting the research and application development of machine learning model and the intelligent transformation of book acquisition. It first collected the Chinese mainland Chinese book subscription catalogue form 2017 to 2022 and the library collection data of Xiamen University Library and processed the data according to the requirements of model construction; Secondly, it conducted feature selection based on factors of book selection, and then performed the data cleaning and standardization of the features; Thirdly, it carried out text features extraction using RoBERTa and with label coding method and expert scoring method it encoded categorical features into numerical type, forming a standardized structured data table; Fourth, it constructed the LightGBM classification model for training and prediction and used the test set data to evaluate the model and analyze the results; Finally, it proposed the model application strategy scheme and follow-up research suggestions. The experimental results show that by utilizing RoBERTa's text understanding ability and LightGBM's efficient classification performance, it can better address the difficulties encountered in the existing intelligent book selection models such as book text feature extraction, high-dimensional discrete features and missing value processing. Ultimately, a Chinese book selection model with good prediction effect can be formed, which can not only effectively utilize bibliographic textual and structured data, but is also practical for application in library acquisition, yielding favorable outcomes. It also provides three model application strategies including recommended collection, selected collection and unsuitable collection to facilitate the practical book selection, thus promoting the research and practical application of book selection machine learning models.

Keywords: University Library; Book Acquisition; Machine Learning Model; RoBERTa; LightGBM