第 45 卷第 3 期
Vol. 45 No. 3 2024

青岛理工大学学报
Journal of Qingdao University of Technology

基于知识图谱的银行中小企业营销模型构建

王秀鸾[1]，张鹏展[1]，杨　鑫[1]，刘　杰[2]
（1. 青岛理工大学 信息与控制工程学院，青岛 266525；2. 青岛银行 数据管理部，青岛 266114）

摘　要：为了提高银行对优质中小企业的识别效率，让数据驱动业务提升，改变银行传统的对公客户营销中存在的信息不对称、以重点大客户为主、过于依赖专家经验等问题，提出了一种基于知识图谱的中小型企业营销模型。通过构建企业间关系网络，形成营销图谱网络，并利用企业图谱特征和评分卡特征筛选优质企业，最后挖掘出潜在的优质企业名单。在相关数据集上的模型训练和实验表明，该模型表现优异，可广泛应用于金融行业客户营销领域。

关键词：客户营销；中小企业营销模型；知识图谱；银行业务；XGBoost 模型
中图分类号：TP181　　**文献标志码**：A　　**文章编号**：1673-4602(2024)03-0162-07

Construction of the marketing model of banks towards small and medium-sized enterprises based on knowledge graph

WANG Xiuluan[1], ZHANG Pengzhan[1], YANG Xin[1], LIU Jie[2]
(1. School of Information and Control Engineering, Qingdao University of Technology, Qingdao 266525, China; 2. Data Management Department, Bank of Qingdao, Qingdao 266114, China)

Abstract: To improve banks' efficiency and data-driven capability in identifying small and medium-sized enterprises (SMEs) and to solve the problems in the traditional corporate marketing such as information asymmetry, too much focus on key clients and over-reliance on expert experience, a marketing model of SMEs based on knowledge graph is proposed. Through the construction of inter-enterprise relationship network, a marketing map network is formed. Then high-quality enterprises are screened by combining the characteristics of the enterprise graph with the score cards, and finally the list of potential high-quality enterprises is excavated. The model training and experiments on relevant data sets show that the model performs well and can be widely used in the client marketing of financial industry.

Key words: client marketing; SME marketing model; knowledge graph; banking business; XGBoost model

近期，各金融机构在中国人民银行印发的《金融科技发展规划(2022—2025 年)》引领下，积极制定数字化转型战略，加快数字化转型的进程。银行传统的对公营销方式主要针对重点企业，而重点企业的筛选主要依赖企业自身财务信息、企业资质及其他相关背书信息，此类企业通常是大中型企业，数据有限、资产有限、风险集中。对于“小而精”、隐形冠军类企业缺乏足够关注。实践证明优质中小企业也可以“聚少成

收稿日期：2023-02-17
基金项目：国家自然科学基金资助项目(42201506)；山东省省级大学生创新训练项目(S202210429055)
作者简介：王秀鸾(1978－)，女，山东潍坊人。硕士，讲师，主要从事大数据、知识挖掘方面的研究。E-mail：1420741063@qq.com。

多”，成为扩大银行利润的重要部分。在大数据技术的加持下，数字化营销浪潮已经奔涌而来。杨宇萍等分析了近年来大数据营销的热点和趋势[1]；詹霞云等从金融机构管理角度设计了对众多小微企业的管理方案[2]；马娴等总结了知识图谱在银行数字化营销的用户经营策略、营销风险监测、营销效果评估等阶段的综合应用[3]；张旭升使用知识图谱构建了零售客户营销与风险领域模型以及配套的分析方法[4]；江震等提出了基于深度学习改进的 VGG-Net16 与 DenseNet 融合人脸识别算法应用于银行客户身份识别过程，不仅降低了运营成本，而且提高了银行客户营销的效率[5]。

本文研究目的是构建商业银行识别对公客户中的优质中小企业的模型，通过大数据分析建模，利用金融知识图谱及机器学习相关技术挖掘“小而精”企业，将“1＋N ”模式转变为“N ＋N ”的营销模式[6]，降低营销成本，促进业务快速发展。创新点在于把知识图谱与企业评分卡结合，将企业的图谱特征纳入企业评分卡维度，并用 XGBoost 构建识别优质中小企业的模型。

1　关键技术介绍

1.1　知识图谱

知识图谱是以图模型的方式描述现实世界中的实体(概念)及实体间关系的知识库[7]。知识图谱通过将业务领域中获取的半结构化、非结构化的数据中的事物或概念转化为实体，并根据业务逻辑建立实体间的网状知识链接，以更清晰的形式描述业务领域复杂的关联逻辑。

构建知识图谱是一项庞杂的工程，涉及本体建模、知识抽取、图存储、知识融合和知识推理等多方面技术[8]。知识抽取主要是识别客观世界中的实体、关系及属性；知识融合是打破多个信息源的限制，通过对不同知识库的整合操作，形成更加统一、全面的知识图谱；知识推理是经过算法推理、分析，挖掘隐藏的知识，从而达到拓展和更新知识库的目的，是构建知识图谱的关键一环。

1.2　XGBoost 模型

XGBoost(Extreme Gradient Boosting)[9]是一种效率更高的梯度提升决策树算法[10]。它与其他统计学模型相比的优势在于，一方面它可以处理“缺失值”，能为缺失的数据设定默认的分支走向，从而减小因数据缺失造成的误差，另一方面它可以同时处理分类和数值特征，使预测模型更加稳定[11]。因此本文选用 XGBoost 作为识别优质企业的基分类器。

XGBoost 算法步骤：

给定数据集 Dataset＝{(x_1, y_1), (x_2, y_2), …, (x_n, y_n)}，其中 $x_n \in R^m$ ，且 $y_n \in \{0,1\}$ ，则

1) 预测模型表示为

$$\widetilde{y}_p = \sum_{m=1}^{m} f_m(x_n), f_m \in F \tag{1}$$

式中：$\widetilde{y}_p$ 为第 p 棵树的预测结果；m 为决策树的数量；x_n 为输入的第 n 个样本；F 为决策树的总和。

2) 目标函数表示为

$$T_t = \sum_{m=1}^{k} \text{loss}[y, \widetilde{y}_P(t-1)_n + f_t(x_n)] + \Omega(f_t) \tag{2}$$

式中：T_t 为第 t 次迭代的目标函数；loss 为损失函数；$\widetilde{y}_p(t-1)_n$ 为前次的预测结果；$f_t(x_n)$为新加入项；$\Omega(f_t)$为正则化项。

化简为

$$T_t' = \sum_{n=1}^{k} [j_n f_n(x_n) + \frac{1}{2} O_n \int_t^2 (x_n) + \Omega(f_t)] \tag{3}$$

式中：j_n 为一阶导数；O_n 为二阶导数。

3) 最优解与目标函数的公式为

$$Q_j^* = -\frac{p_n}{s_n + \lambda} \tag{4}$$

$$T_t^{'}=-\frac{1}{2}\sum_{n=1}^{k}\frac{p_j^2}{s_n+\lambda}+\lambda T \tag{5}$$

式中：Q_j^* 为最优解；$T_t^{'}$ 为目标函数的值。

2 营销原理及构建过程

2.1 模型原理

本文研究的中小型企业营销模型是一种基于知识图谱的网络模型，如图 1 所示，通过构建企业间的关联关系形成知识图谱，并优化图谱网络结构形成营销图谱，选取企业的特征维度并计算特征值，使用 XGBoost 算法获得各子图中的核心企业，通过企业间的交易强度等信息挖掘潜在优质企业。

2.2 基于知识图谱的营销模型构建

模型构建分为三步：基础知识图谱网络构建、营销图谱网络构建、模型训练及挖掘，如图 2 所示。

第一步，构建基础知识图谱网络。本项目构建的是金融行业的知识图谱，故采用自顶向下的方式构建图谱本体模型[12]。为充分挖掘企业间关系，在数据选取上不仅使用了商业银行对公客户数据，还融合行外企业工商、股权、政务等数据。利用知识图谱的构建工具进行数据清洗、知识抽取、知识融合、知识加工，最终将结构化的二维数据转化为图结构的数据，形成基础知识图谱网络。

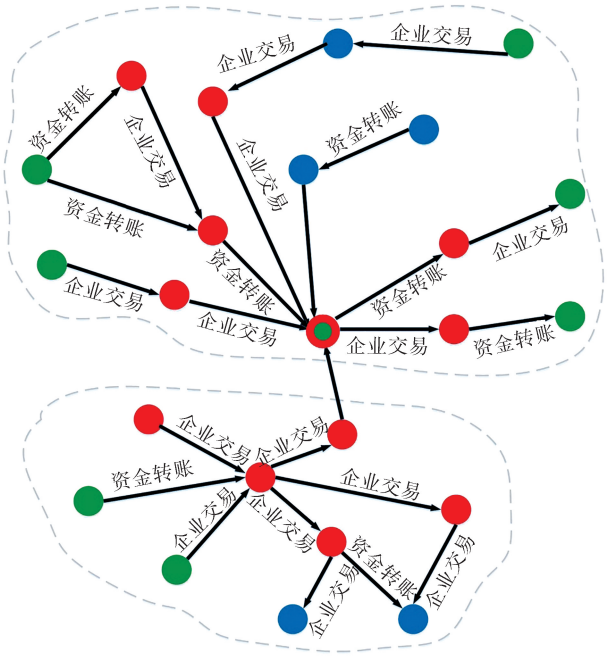

图 1 基于知识图谱的中小型企业营销模型

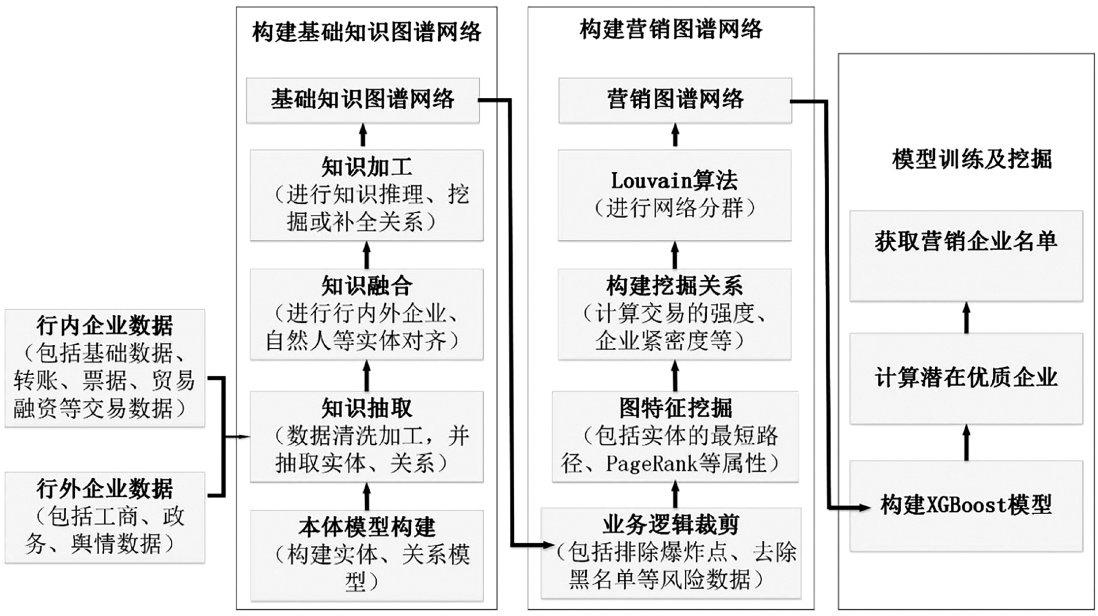

图 2 营销模型的构建路线

第二步，构建营销图谱网络。通过对基础知识图谱进行优化挖掘，构建营销图谱网络，具体操作如下：

1）排查爆炸点。把网络中含有 1000 条及以上关系的企业节点称为爆炸点，爆炸点影响营销效果及图谱展示。

2）去除风险企业。黑名单及疑似黑名单企业不能纳入营销范围之内，企业黑名单来源于商业银行黑名单数据，疑似黑名单是根据黑名单同地址、交易频率等规则挖掘而来。

3）计算企业特征。通过图挖掘 Pregel 算法，计算企业最短路径、PageRank 等属性；通过已有的交易数据，计算企业交易的强度、企业紧密度等特征，相关计算公式如下：

$$X=100\times e^{\frac{-\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}}{\bar{x}}} \tag{6}$$

式中：x_i 为企业相邻交易间隔天数；$\bar{x}$ 为相邻转账间隔天数平均值；X 为交易时间稳定性。

$$Y=100\times e^{\frac{-\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}{\bar{y}}} \tag{7}$$

式中：y_i 为企业一次交易；$\bar{y}$ 为月平均交易次数；Y 为交易次数稳定性。

$$Z=\frac{1}{1+e^{\frac{\sum_{i=1}^{n}z_i/\sum_{i=1}^{n}y_i}{z}}} \tag{8}$$

式中：Z 为交易重要性；z_i 为每次交易金额。

$$R=0.4X+0.2Y+0.4Z \tag{9}$$

式中：R 为企业关联紧密度，选取 $R\geqslant 0.4$ 的交易纳入营销网络。

4）使用 Louvain 算法[13]分群。由于原始网络数据量大且网络稀疏，为了便于寻找核心企业，选用 Louvain 社区发现算法将形态特征类似的节点归类，最大化知识网络的模块度[14]。

第三步，模型训练及挖掘。基于构建好的营销图谱网络，选取企业特征进行模型训练并挖掘，输出营销名单。特征维度直接影响模型效果，因此选取特征不仅结合企业评分卡加入了企业基本信息、交易特征、风险特征，还融入了企业网络特征。最终，为每个企业建立了 24 维特征，如图 3 所示。

图 3　某企业 24 维特征

判断是否优质企业本质上是一个分类问题，XGBoost 模型是解决分类问题最好的模型之一。利用某商业银行认定的核心企业及部分价值户作为优质企业样本，另选取部分企业作为非优质企业样本，使用 XGBoost 进行模型训练，获取优质企业，并将优质企业关联较紧密的客户列为营销名单。

3　模型实现及结果

3.1　营销图谱实现

本文研究的营销模型是在某商业银行智能图仓系统上实现的，原始数据存储在 Hive 数据库中，数据的处理加工通过 Spark 任务实现，图数据采用 Arango 数据库存储。使用知识图谱构建工具完成了基础知识图谱网络构建，涉及企业股权、担保、高管、实控人、转账和票据等 20 余种关系，属性信息 200 余个，企

业实体数量约 2 亿个，关系数量约 10 亿条，如图 4 所示。

经过优化图谱网络，排除爆炸点 200 余个，去除风险企业约 1 万个，经过 Louvain 算法模块化，最终完成营销网络构建，如图 5 所示。

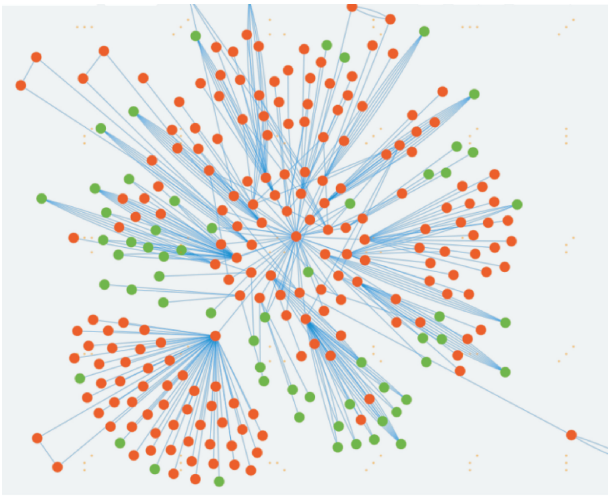

图 4 某企业的三度展开关系图谱

● 本行客户；● 他行客户

3.2 XGBoost 模型建立及训练

经营销网络构建及特征处理，结合某银行优质企业数据，构造用于训练模型的数据样本及总样本。总样本数量为 2 万个，每个样本有 25 个变量，其中 24 个为解释变量，另外 1 个变量 flag 用来记录企业类别，核心企业的 flag 值为 1，非核心企业 flag 值为 0。选取总样本中的 3/4 作为训练集，其余作为测试集。

模型在训练集上训练时，随着训练次数增加，其损失逐渐接近 0，AUC（Area Under ROC Curve）达到了 0.9635，这代表模型很好地拟合了训练集数据，但利用测试集数据进行测试时 AUC 为 0.6030，准确率较低，说明模型训练的过拟合了，因此加入正则化项进行约束。通过不断调整正则化系数，发现当正则化系数 reg_alpha 为 0，正则化系数 reg_lambda 为 2 时，模型是最优的，预测得到的 AUC 最大为 0.8020。最终确定 XGBoost 模型的最优参数见表 1。

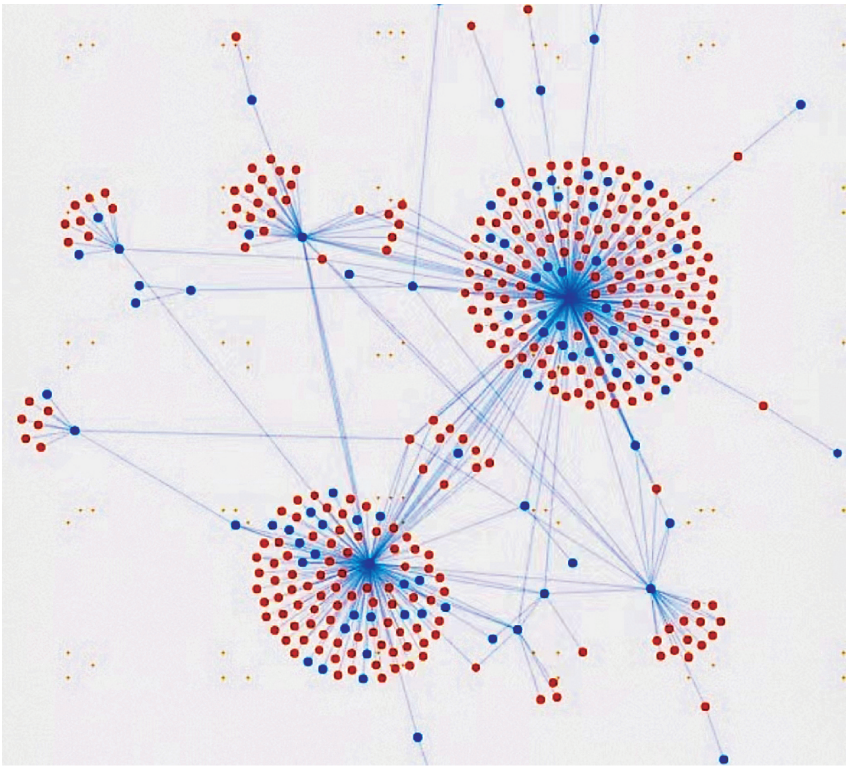

图 5 经Louvain算法处理后的营销图谱

● 本行客户；● 他行客户

3.3　模型评价指标

在使用机器学习算法的过程中，针对不同业务场景需要选用不同的模型评价指标。评价一个二值分类器的优劣最常用的是 ROC(Receiver Operating Characteristic)曲线和 AUC 值[15]。

ROC 曲线，横轴用来表示假正例率(False Positive Rate，FPR)，即表示误判为正的比例；纵轴用来表示真正例率(True Positive Rate，TPR)，也就是正确判断为正的比例。从原点处开始，如果一个正样本被分类器正确预测为正，则 TPR 增加；若分类器将一个负样本误判为正，进而 FPR 增加。

AUC 值是指 ROC 曲线与 x 轴围成的面积，本文用 AUC 值来判断分类器的性能，容易看出，AUC 是一个小于 1 的正数，且 AUC 统计量越大，模型的分类效果越好。

表 1　XGBoost 模型最优参数

参数	最优值
max_depth	6
learning_rate	0.8
n_estimators	210
objective	binary:logistic
subsample	1
n_jobs	16
reg_alpha	0
reg_lambda	2
gamma	0.5
colsample_bytree	0.85

3.4　XGBoost 模型结果与分析

经过对 XGBoost 模型训练 210 次，AUC 值最开始为 0.6200，最后达到 0.8293，如图 6 所示。从图 6 中还可以看出，在训练前期 AUC 值的增加比较明显，基本在训练迭代次数达到 200 次以后，AUC 值增加趋势变缓，趋于稳定。

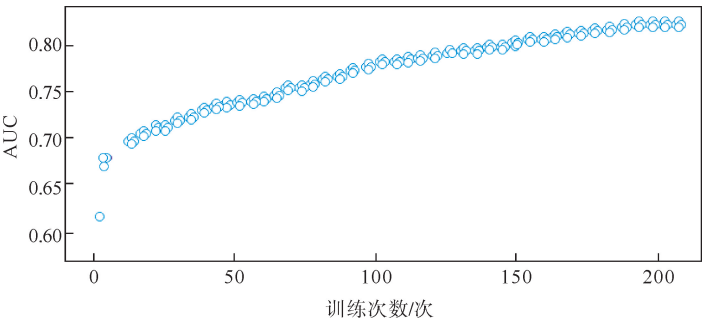

图 6　训练效果

导入测试集并在训练模型上进行测试，根据测试值得到 ROC 曲线和 AUC 值，如图 7 所示。

将训练好的模型用于识别营销图谱中的优质企业，新挖掘出优质企业约 100 个，挖掘潜在营销名单企业 200 余个。

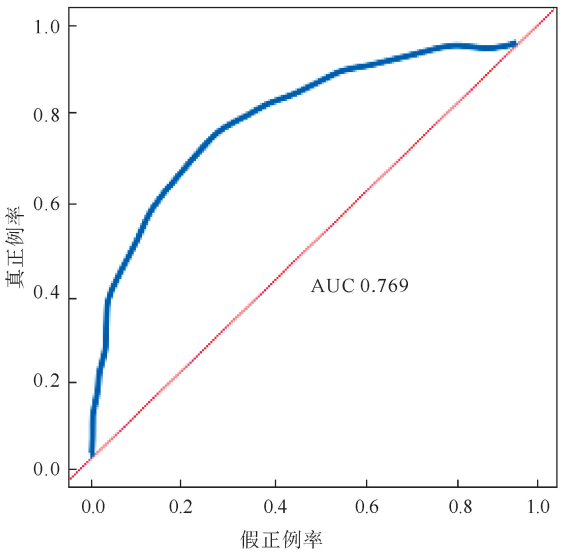

图 7　ROC 曲线和 AUC 值

4　结束语

本文通过分析某银行对公客户营销的痛点，利用知识图谱及机器学习等技术构建中小企业营销模型，挖掘优质的中小企业，极大地提高了银行的营销效率，扩大了银行的利润来源，具有很高的实际应用价值。本文最终得到的营销模型还有可优化空间，目前图谱构建使用的企业数据有限，可以通过增加企业产业链等数据丰富图谱关系，另外通过增加企业特征维度可能对模型挖掘效果有所提升。

知识图谱是企业数字化转型不可或缺的工具，今后将在知识图谱领域不断探索，将知识图谱应用到越来越多的场景中，为银行业务提供更好的智能化服务，不断提升中小商业银行智能化水平。

参考文献(References):

[1] 杨宇萍，陈章旺. 大数据营销的研究热点及趋势:基于知识图谱的量化研究[J]. 商业经济研究，2020(3):87-89.
YANG Yuping，CHEN Zhangwang. Research hotspot and trend of big data marketing: Quantitative research based on knowledge graph[J]. Journal of Commercial Economics,2020(3):87-89.

[2] 詹霞云，冀用武. 基于商业银行网点对公长尾客户管理的思考[J]. 时代金融,2020(9):37-39.
ZHAN Xiayun，JI Yongwu. Thoughts on management of corporate long tail clients based on commercial bank outlets[J]. Times Finance，2020(9):37-39.

[3] 马娴，汪昀，梅影. 知识图谱在银行数字化营销中的应用[J]. 银行家，2020(1):128-129.
MA Xian，WANG Yun，MEI Ying. Application of knowledge graph in digital marketing of banks[J]. The Chinese Banker，2020(1):128-129.

[4] 张旭升. 面向兰州银行零售客户的知识图谱设计与研究[D]. 兰州:兰州大学，2021.
ZHANG Xusheng. Design and research of knowledge graph for retail clients of the bank of Lanzhou[D]. Lanzhou: Lanzhou University，2021.

[5] 江震，曲娜，胡从强. 基于深度学习的银行客户身份识别算法研究[J]. 青岛理工大学学报，2023，44(1)：147-152.
GANG Zhen，QU Na，HU Congqiang. Research on bank customer identification system based on deep learning[J]. Journal of Qingdao University of Technology，2023，44(1)：147-152.

[6] 谭建，闫丽娜，李萌. 物流导向型供应链金融模式构建[J]. 商业经济研究，2020(1):157-159.
TAN Jian，YAN Lina，LI Meng. Construction of logistics oriented supply chain financial model[J]. Journal of Commercial Economics，2020(1):157-159.

[7] 王鑫，邹磊，王朝坤，等. 知识图谱数据管理研究综述[J]. 软件学报，2019，30(7):2139-2174.
WANG Xin，ZOU Lei，WANG Chaokun，et al. Research on knowledge graph data management: A survey[J]. Journal of Software，2019，30(7):2139-2174.

[8] 王昊奋，丁军，胡芳槐，等. 大规模企业级知识图谱实践综述[J]. 计算机工程，2020，46(7):1-13.
WANG Haofen，DING Jun，HU Fanghuai，et al. Survey on large scale enterprise-level knowledge graph practices[J]. Computer Engineering，2020，46(7):1-13.

[9] 林奕晨，周鹏，潘悦，等. 荆州市洪涝灾害影响因子探究及风险评估:基于随机森林和 XGBoost 算法[J]. 中国农村水利水电，2022(6):125-132.
LIN Yichen，ZHOU Peng，PAN Yue，et al. Influencing factor research and risk assessment of flood disasters in Jingzhou City: Based on random forest and XGBoost algorithm[J]. China Rural Water and Hydropower，2022(6):125-132.

[10] 李怡静，孙晓敏，郭玉银，等. 基于梯度提升决策树算法的鄱阳湖水环境参数遥感反演[J]. 航天返回与遥感，2020，41(6):90-102.
LI Yijing，SUN Xiaomin，GUO Yuyin，et al. Remote sensing retrieval of water quality parameters in Poyang Lake based on the gradient boosting decision tree algorithm[J]. Spacecraft Recovery & Remote Sensing，2020，41(6):90-102.

[11] 王晓晖，张亮，李俊清，等. 基于遗传算法与随机森林的 XGBoost 改进方法研究[J]. 计算机科学，2020，47(S2):454-458.
WANG Xiaohui，ZHANG Liang，LI Junqing，et al. Study on XGBoost improved method based on genetic algorithm and random forest[J]. Computer Science，2020，47(S2):454-458.

[12] 袁俊，刘国柱，梁宏涛，等. 知识图谱在商业银行风控领域的研究与应用综述[J]. 计算机工程与应用，2022，58(19):37-52.
YUAN Jun，LIU Guozhu，LIANG Hongtao，et al. Summary of research and application of knowledge graphs in risk management field of commercial banks[J]. Computer Engineering and Applications，2022，58(19):37-52.

[13] 张微薇，刘盾，贾修一. 基于 XGBoost 的三分类优惠券预测方法[J]. 南京航空航天大学学报，2019，51(5):643-651.
ZHANG Weiwei，LIU Dun，JIA Xiuyi. Three classified coupon prediction based on XGBoost algorithm[J]. Journal of Nanjing University of Aeronautics & Astronautics，2019，51(5):643-651.

[14] TRAN Q M，NGUYEN H D，HUYNH T，et al. Measuring the influence and amplification of users on social network with unsupervised behaviors learning and efficient interaction-based knowledge graph[J]. Journal of Combinatorial Optimization，2021，44(4):1-27.

[15] 李子言. 大数据背景下 ROC 曲线介绍与应用[J]. 科教导刊，2021(14):81-84.
LI Ziyan. Introduction and application of ROC curve under the background of bif data[J]. Disciplines Exploration，2021(14):81-84.

(责任编辑 姜锡方;英文校审 程文华)