王炜辰. 基于机器学习的在线快时尚商品退货预测研究[J]. 智能计算机与应用,2024,14(9):88-92. DOI:10.20169/j.issn.2095-2163.240913

基于机器学习的在线快时尚商品退货预测研究

王炜辰
(南京审计大学 商学院, 南京 211800)

摘 要: 商品退货是影响在线快时尚运营绩效的重要因素,有效识别退货的重要因素、预测退货行为对提高在线快时尚业的运营绩效有重要意义。本研究基于真实在线快时尚商品退货数据,对比了 7 个应用机器学习模型的预测表现,包括决策树、随机森林、梯度提升决策树、轻量级梯度提升机、极端梯度提升和分类特征支持的梯度提升 6 个基础模型和一个堆叠模型,并对 6 个基础预测模型中的特征重要性进行比较,以识别影响退货的重要变量。通过 6 个评价指标来评估 7 个模型的退货预测的表现,结果表明影响产品退货的最重要的 3 个因素,即订单总花销、推荐购买价格和付款方式;分类特征支持的梯度提升模型的综合预测表现优于其他 5 个基础模型,更加适用于识别在线快时尚商品的退货预测;而 Stacking 组合堆叠模型在该数据集中并没有进一步提高预测的精度。

关键词: 快时尚; 退货预测; 机器学习; 组合模型
中图分类号: TP181;F713.36 **文献标志码**: A **文章编号**: 2095-2163(2024)09-0088-05

Research on online fast fashion product return prediction based on machine learning

WANG Weichen
(School of Business, Nanjing Audit University, Nanjing 211800, China)

Abstract: Merchandise returns are an important factor that affects the operational performance of online fast fashion. Effectively identifying important factors affecting returns and predicting return behavior are important to improve the operational performance of the online fast fashion industry. Based on real online fast fashion merchandise return data, this study applies machine learning models to compare the predictive performance of seven models, including this study compares the prediction performance of seven models, including six base models of Decision Tree, Random Forest, Gradient Boosting Decision Tree, LightGBM, XGBoost and CatBoost and a stacking model, and compares the importance of features in the six base prediction models to identify the important variables. The performance of the seven models for return prediction is evaluated by six evaluation metrics. The results show that the three most important factors affecting product returns, i.e., total order spend, recommended purchase price, and payment method; the combined prediction performance of the CatBoost model outperforms the other five base models and is more suitable for identifying return predictions for online fast fashion items; and the Stacking combination model does not further improve the prediction accuracy in this dataset.

Key words: fast fashion; return forecast; machine learning; combination model

0 引 言

随着消费者需求的不断升级,进而推动了全球服装零售业规模的不断壮大。在人们实现购物简单化和多样化的同时,也给零售商增加了多样化的收入渠道。而电子商务退货是网上购物过程中一个正常的、不可预期的部分。如果网站上的商品与实际收到的商品不符,消费者会根据商家的退货策略进行退货[1]。美国零售协会(NRF)估计,电商产品的退货率比实体店高两到三倍。对于专门从事快时尚的在线零售商来说,通常有超过 50%的商品购买后会被顾客退货,大量的退货给在线零售商带来额外的运输、整理和检查成本,对其经营绩效产生巨大影响。

作者简介: 王炜辰(2000-),女,硕士研究生,主要研究方向:大数据驱动的商务分析与决策优化。Email:1136218161@qq.com

收稿日期: 2023-05-18

成功而实用的策略应该侧重于在交易发生之前识别具有极高退货率的商品[2]。Hess 等[3]是最早研究商品退货的，并为服装零售商提供了预测退货时间的模型；Ma 等[4]提出了自回归类型(Autoregressive-type)的统计模型来预测退货数量和时间；Dzyabura 等[5]提出梯度增强回归树模型，利用图像处理技术来精确预测产品的回报率；Tüylü 等[6]专注于客户偏好的产品的退货问题，采用 EML 方法中的堆叠和投票模型更准确地预测产品退货。

本文基于一个在线快时尚零售商两年的商品交易和退货大数据，运用多种机器学习模型，分析对退货预测产生影响的重要特征变量，并对比各个机器学习模型的预测精度；从订单属性(如：购买方式、购买总数量)、商品属性(如：颜色、尺寸)和顾客属性(如：顾客平均消费、顾客购买数量)3 个角度构建预测特征变量，选取了决策树(DT)、随机森林(RF)、梯度提升决策树(GBDT)、轻量级梯度提升机(LightGBM)、极端梯度提升(XGBoost)以及分类特征支持的梯度提升(CatBoost)6 个基础模型和一个堆叠(Stacking)模型进行分析和预测。为了对预测的精度进行严格的评估，本文选取了多个绩效指标，包括准确性、精确率、召回率、*F*1 值、*ROC* 曲线下面积(*AUC*)和精确召回面积(*AUCPR*)指数。本文的研究有助于更好地理解影响快时尚零售业中顾客退货行为，通过准确预测退货为企业的退货策略改进和退货运营调度计划提供依据。

1　模型框架

本文选取了 6 个基础模型即 DT、RF、GBDT、LightGBM、XGBoost 和 CatBoost，通过贝叶斯优化方法对 6 个基础模型进行调参，运用五折交叉验证方法建立 Stacking 组合预测模型，最终通过计算 6 个性能指标即准确率、精确率、召回率、*F*1 值、*AUC* 及 *AUCPR* 对 7 个模型进行比较。

1.1　预测因子

本文从原始数据中提取出了 22 个预测变量，可分为 3 个类别，包括订单属性、商品属性和客户属性。在订单属性中包括订单花销、购买方式等；产品属性中包括商品尺寸、商品颜色等，两者都包含了 9 个预测变量；顾客属性中包括顾客购买数量、顾客平均花销、顾客购买频次和顾客花销 4 个预测变量；产品退货预测的特征变量详细描述见表 1。

表 1　产品退货预测的特征变量

Table 1　Characteristic variables for product return forecasting

类别	属性	特征数量
订单属性	购买方式	10
	设备编号	5
	是否使用优惠卷	2
	购买星期	1
	购买月份	1
	订单购买数量	1
	订单折扣	1
	优惠卷价值	1
	订单花销	1
商品属性	商品颜色	642
	商品产品组	32
	商品尺寸	29
	颜色流行度	3
	推荐购买价格	1
	商品总价值	1
	商品单价	1
	商品折扣	1
	商品购买数量	1
顾客属性	顾客平均花销	1
	顾客购买频率	1
	顾客花销	1
	顾客购买数量	1

1.2　预测模型

本研究选取 7 个机器学习模型对快时尚在线零售的退货预测问题进行研究，包括 6 个基本预测模型和一个组合预测模型。

(1)决策树(DT)通过建立一个二叉树对数据进行分类的概率分析模型[7]；

(2)随机森林(RF)采用装袋法将多个基础分类器与均匀分布的权重相结合，以提高单个决策树的预测性能[8]；随机森林是树预测器的集成，每棵树都依赖于独立采样的随机向量的值，所有树都有一样的分布[9]；

(3)梯度提升决策树(Gradient Boosting Decision Tree，GBDT)最早由 Friedman[10]在 2001 年提出，采用加法模型，以决策回归树为基学习器进行线性组合与前向分布计算，迭代训练而构成的一种集成学习模型；

(4)LightGBM、XGBoost 和 CatBoost 模型

LightGBM、XGBoost 和 CatBoost 模型都是基于梯度提升决策树模型的开源机器学习模型，被广泛

应用于分类和回归问题,都采用类似的迭代决策树生成方法,通过多次迭代不断优化模型的性能[11]。

LightGBM 模型是由微软开发的一种快速高效的 GBDT 模型,具有很高的训练和预测速度,并且能够处理大规模数据集,采用了一些独特的策略来降低内存消耗,提高训练速度,例如直方图算法策略、带深度限制的叶子分裂策略和互斥特征捆绑策略等。

XGBoost 模型是由 Chen 等[12]在 2014 年创立的 GBDT 模型,也是目前最流行的 GBDT 模型之一,支持多种语言实现,包括 Python, R, C++等,并且具有很高的可扩展性和鲁棒性。XGBoost 采用了一些独特的算法来提高模型的性能,例如贪心算法、正则化项算法等。

CatBoost 模型是由 Yandex 公司[13]开发的基于 GBDT 模型的机器学习模型,采用排序提升算法(Ordered Boosting),有效减少了过拟合、梯度偏差和预测偏移等问题,从而提升训练和预测速度。CatBoost 还能自动处理类别特征或文本特征,并自动进行独热编码,能够有效地管理和利用内存,而无需显著增加系统资源。

(5)Stacking 模型

Stacking 模型的基本思路是预测多个基模型,然后将基模型的输出数据合并到元模型中,然后用元模型进行预测,输出预测结果[14]。Stacking 模型主要有两层,第一层由多个基模型组成,该层每一个基模型使用的数据均是来自原始的数据集;第二层则为元模型,通常选择较为简单的模型,元模型中训练数据是来自第一层每一个基模型对原始训练数据进行 k 折交叉验证得到的预测数据拼接而成的数据集,测试数据是每一个基模型对原始的测试集进行预测得到的预测结果取均值并拼接而成的数据集[15]。Stacking 模型目的是将强大而多样的学习者群体融合在一起,本文利用 6 个基础模型作为基模型,逻辑回归作为元模型,对模型进行融合。

本文采取贝叶斯优化方法对各个基模型的超参数进行选择。贝叶斯优化是一种通过构建目标函数的后验分布来进行优化的方法,其基本思想是将超参数看作概率分布中的随机变量,并通过已知观察结果对该分布进行更新,从而缩小猜测范围,找到最优解。本文将训练集的样本按照 80% 和 20% 的比例进行进一步分割,形成训练集和验证集,并通过验证集进行超参数的优化。

2 实验

实验评估的目的是探究影响该在线快时尚零售业的商品退货的重要因素,以及探究最适用于预测退货的模型。为了实现这一目标,首先计算出 6 个基本模型中每个预测因子的特征重要性;其次,使用 6 个性能指标来评估 6 个基本模型和一个 Stacking 组合模型获得的性能。

2.1 数据样本

本文的实验数据来源于德国一家在线快时尚零售商从 2014 年 1 月至 2015 年 12 月的历史交易和退货记录。在原数据集中,2014 年至 2015 年一共有 2 666 263 条商品交易数据。把“购买数量大于零”以及“购买数量大于等于退货数量”作为筛选条件,过滤了少量无效样本数据,剩余数据共有 2 627 927 条。本文将 2014 年 1 月至 2015 年 9 月的数据作为训练集,2015 年 10 月至 2015 年 12 月的数据作为测试集进行实验。

2.2 特征重要性

特征重要性是一个衡量每个输入特征对模型预测结果贡献的指标,即某个特征上的微小变化如何改变预测结果。利用 6 个模型对数据进行了自变量重要性分析,特征重要性分布的箱形图如图 1 所示。

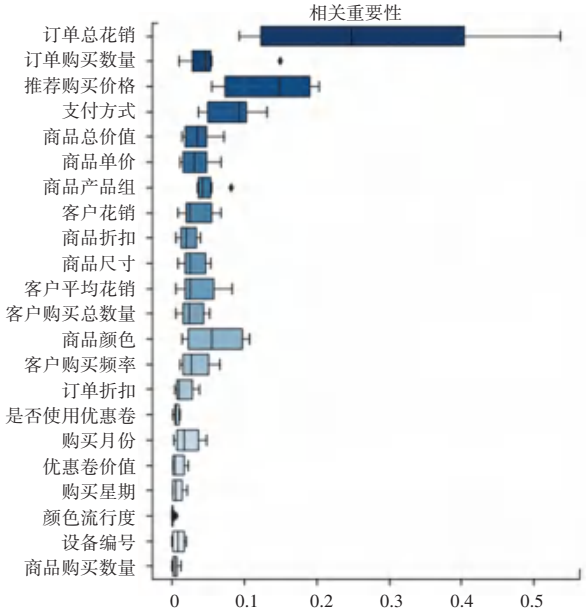

图 1 6 个基础模型的特征重要性分布的箱形图

Fig. 1 Box plot of feature importance distribution for the six base models

综合所建立的 6 个基础模型的特征重要性的排序得出,订单花销、推荐购买价格和付款方式是最重要的影响因子;较重要的是商品的颜色和订单的购

买数量。由此表明，订单属性和商品属性对该数据集的预测贡献率相对较大，为影响该快时尚商品退货率的影响因素。

2.3　模型性能评价

本文将 7 个分类模型进行对比实验，利用混淆矩阵得到准确率（*Accuracy*）、精确率（*Precision*）、召回率（*Recall*）、*F*1 值（*F*1 - Score）、*AUC* 以及 *AUCPR* 6 个指标值，基础模型和堆叠模型在训练集和测试集上的性能比较的结果如图 2 所示。

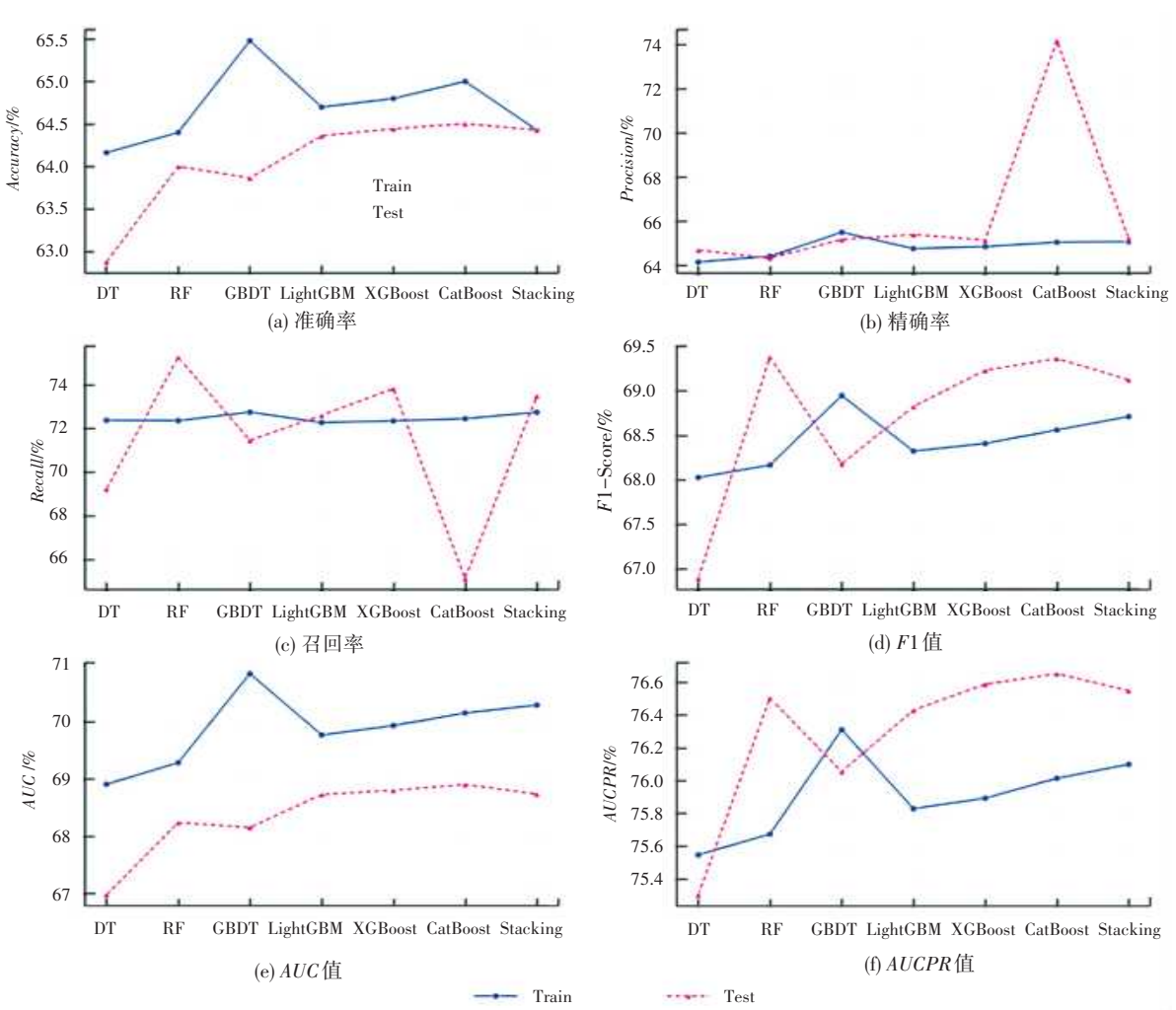

图 2　基础模型和堆叠模型在训练集和测试集上的性能比较

Fig. 2　Performance comparison of the base models and the Stacking model on the training and testing sets

综合 7 个模型在 6 个性能指标上的预测表现得出：即使训练集中 GBDT 模型的性能表现为最优，但是在测试集中，Catboost 模型优于其他 6 个分类模型。分别观测 6 个性能指标的折线图，在准确率、精确率、*AUC* 值和 *AUCPR* 值上的 CatBoost 模型是性能表现均优于其他模型，特别是在精确率上表现尤为突出；LightGBM、XGBoost 和 Stacking 模型的综合预测能力基本持平；决策树模型最差；在准确率、*F*1 值、*AUC* 值和 *AUCPR* 值的折线图中，7 个模型的预测性能趋势较为相似；在召回率的折线图中，随机森林模型表现最好。

通过比较 6 个基础预测模型和一个组合预测模型的 6 个性能指标，综合预测结果得出 CatBoost 模型具有较强的泛化能力和较高的预测精度，最适用于预测该服装销售商店的退货预测。

3　结束语

准确预测客户是否退货且分析客户退货的关键影响因素能有效提高客户服务质量和效率。本文建立了 7 个机器学习算法模型来预测德国一家在线快时尚零售商店的商品退货概率，主要目的是探究影响快时尚零售业商品退货的重要变量；建立 6 个基础模型及一个堆叠模型，通过 6 个评判指标识别最优的预测模型。

研究表明,对于该快时尚商品退货的影响因素而言,最重要的 3 个因素为订单总花销、商品价格和付款方式。在所构建的 6 个基础模型中,Catboost 模型的综合预测表现最优,具有较强的泛化能力和较高的预测精度,最适用于预测该时装销售商店的退货预测;Satcking 组合预测模型并没有进一步提高预测精度。

参考文献

[1] HUANG X, CHOI S M, CHING W K, et al. On supply chain coordination for false failure returns: A quantity discount contract approach[J]. International Journal of Production Economics, 2011, 133(2): 634-644.

[2] BONIFIELD C, COLE C, SCHULTZ R L. Product returns on the Internet: A case of mixed signals? [J]. Journal of Business Research, 2010, 63(9-10):1058-1065.

[3] HESS J D, MAYHEW G E. Modeling merchandise returns in direct marketing[J]. Journal of Interactive Marketing, 1997, 11(2): 20-35.

[4] MA J, KIM H M. Predictive model selection for forecasting product returns[J]. Journal of Mechanical Design, 2016, 138(5): 054501.

[5] DZYABURA D, EL KIHAL S, IBRAGIMOV M. Leveraging the power of images in predicting product return rates[J]. SSRN Electronic Journal, 2018(5): 1-33.

[6] TÜYLÜ A N A, EROGLU E. Prediction of product return rates with ensemble machine learning algorithms[J]. Journal of Engineering Research, 2022. DOI:10.36909/jer.13725

[7] 栾丽华, 吉根林. 决策树分类技术研究[J]. 计算机工程,2004(9):94-96.

[8] WITTEN I H, FRANK E, HALL M A. Data Mining: Practical Machine Learning Tools and Techniques[M]. St. Louis:Morgan Kaufmann,2011.

[9] CUTLER A, CUTLER D R. Random forests[J]. Machine Learning, 2004, 45(1):157-176.

[10] FRIEDMAN J H. Greedy function approximation: A gradient boosting machine[J]. Annals of Statistics,2001,29(5): 1189-1232.

[11] HASTIE T, TIBSHIRANIR, FRIEDMAN J H. Elements of Statistical Learning[J]. Technometrics, 2009, 45(3):267-268. DOI:10.1198/tech.2003.s770

[12] CHEN T, GUESTRIN C. Xgboost: A scalable tree boosting system[C]//Proceedings of the 22nd ACM Sigkdd International Conference on Knowledge Discovery and Data Mining. IEEE, 2016: 785-794.

[13] PROKHORENKOVA L, GUSEV G, VOROBEV A, et al. CatBoost: Unbiased boosting with categorical features[J]. arXiv preprint arXiv:1706.09516, 2017. DOI:10.48550/arXiv.1706.09516

[14] ZENKO B, TODOROVSKI L, DZEROSKI S. A comparison of stacking with meta decision trees to bagging, boosting, and stacking with other methods[C]//Proceedings of IEEE International Conference on Data Mining. IEEE, 2001: 669-670.

[15] 杜帅帅. 基于 Stacking 集成学习的贷款违约预测模型研究[D]. 大连: 东北财经大学,2022.