Modern Management 现代管理, 2026, 16(5), 328-337
Published Online May 2026 in Hans. https://www.hanspub.org/journal/mm
https://doi.org/10.12677/mm.2026.165107

Hans汉斯

机器学习在跨境电商海关数据评价中的应用研究

章舒晗，孔晓丹

上海海关学院海关与公共管理学院，上海

收稿日期：2026年4月8日；录用日期：2026年4月20日；发布日期：2026年5月27日

摘　要

海关数据评价是跨境贸易监管体系中的关键环节，其核心在于从高维、多源及异构的申报数据中识别潜在风险。随着跨境电商等新型贸易业态的快速发展，传统基于人工经验与规则库的方法在处理复杂数据时存在着效率低下与适应性不足的问题。机器学习作为人工智能的重要分支，依托统计建模与数据驱动机制，为海关数据评价提供了新的技术路径。本文基于机器学习与统计学交叉融合的视角，以跨境电商海关风险预警为应用场景，构建“专家评分–机器学习预测”的混合评价方法。通过引入岭回归与多输出回归模型，实现海关数据定性指标的定量化处理，完成14项定性风险指标的自动化预测。研究结果表明，该方法在提升海关数据处理效率与评价结果一致性方面具有一定优势，可为海关风险识别与智能监管提供切实的方法支撑。

关键词

机器学习，海关数据评价，跨境电商，风险预警，多输出回归

Research on the Application of Machine Learning in Customs Data Evaluation in Cross-Border E-Commerce

Shuhan Zhang, Xiaodan Kong

School of Customs and Public Administration, Shanghai Customs University, Shanghai

Received: April 8, 2026; accepted: April 20, 2026; published: May 27, 2026

Abstract

Customs data evaluation is a crucial link in the cross-border trade supervision system, its core being

文章引用: 章舒晗, 孔晓丹. 机器学习在跨境电商海关数据评价中的应用研究[J]. 现代管理, 2026, 16(5): 328-337.
DOI: 10.12677/mm.2026.165107

the identification of potential risks from high-dimensional, multi-source, and heterogeneous declaration data. With the rapid development of new trade formats such as cross-border e-commerce, traditional methods based on human experience and rule bases suffer from inefficiency and insufficient adaptability when processing complex data. Machine learning, as an important branch of artificial intelligence, provides a new technical path for customs data evaluation by relying on statistical modeling and data-driven mechanisms. This paper, based on the cross-border integration of machine learning and statistics, takes cross-border e-commerce customs risk early warning as an application scenario and constructs a hybrid evaluation method of "expert scoring-machine learning prediction". By introducing ridge regression and multi-output regression models, the qualitative indicators of customs data are quantified, and the automated prediction of 14 qualitative risk indicators is achieved. The results show that this method has certain advantages in improving the efficiency of customs data processing and the consistency of evaluation results, and can provide practical methodological support for customs risk identification and intelligent supervision.

Keywords

Machine Learning, Customs Data Evaluation, Cross-Border E-Commerce, Risk Warning, Multi-Output Regression

1. 引言

海关数据评价是实现进出口监管、风险防控与贸易便利化的重要技术手段，其主要目标在于通过对申报数据的系统分析，识别潜在的税收风险、安全风险与合规风险。近年来，随着跨境电商的迅猛发展，海关业务数据呈现出规模持续扩大、结构日趋复杂及类型高度多样化的特征[1]。在此背景下，传统依赖人工审核与规则匹配的评价模式存在以下不足：一是对人工经验依赖较强，主观性较高；二是规则库更新滞后，难以应对新型风险模式；三是在高维数据环境下缺乏有效的特征提取与模式识别能力。亟需引入数据驱动的方法以提升海关数据评价的科学性与自动化水平。

机器学习通过对历史数据的学习构建预测模型，能够实现对未知样本的自动分类与风险评估，在金融风控、智能制造等领域已得到广泛应用[2]。将其引入海关数据评价，有助于提升风险识别的准确性和数据处理效率。现有研究多关注于直接构建风险分类或预测模型，而对于“专家经验如何结构化表达并实现自动化复现”的问题关注较少。如何将专家对定性指标的判断逻辑转化为可学习的模型，是连接传统经验方法与数据驱动方法的重要桥梁。本文并不以“预测真实风险”为直接目标，而是以“模拟专家评分机制”为切入点，尝试探索定性指标风险量化的可行路径。

2. 机器学习的理论基础

机器学习(Machine Learning, ML)就是让计算机从数据中进行自动学习，得到某种知识或规律。其核心在于利用数据训练模型，使模型能够对新数据做出预测或决策。机器学习是一门研究数据学习和预测的多领域交叉学科，在科学研究、生产生活中有着广泛应用[3]。机器学习作为人工智能的一个关键子领域[4]，其利用统计学方法和模型，使机器自动学习基于数据的模型。机器学习在过去几十年取得了飞速发展，并对各个领域产生了深远影响[5]。

3. 机器学习在海关数据评价中的应用

3.1. 数据来源与研究场景

本研究聚焦跨境电商进口环节的海关风险预警场景，该业务情境涉及海量申报数据的实时处理与潜在风险模式识别，对传统依赖人工经验的审核方式提出了严峻挑战。研究采用 GitHub 平台发布的公开数据集作为实证分析基础，该数据集包含 54,000 条人工生成的模拟交易记录，涵盖 22 个关键属性变量[6]，并经由条件生成对抗网络(CTGAN)进行数据合成与增强[7]，在保持与源数据分布特征一致的前提下，有效扩充了样本规模并规避了商业隐私泄露风险。合成数据可以通过针对性的数据补充和强化，解决数据匮乏、数据质量不足等问题；可以规避数据隐私、安全、保密等风险，在医疗、金融等领域意义重大；还可以模拟和生成现实世界中难以采集到的边缘场景，保持数据的多样性[8]。

基于跨境电商监管的特定需求，参照贸易手册对进口类型(import type)编码为 15 的交易记录进行针对性筛选，最终获得 954 条符合研究目标的跨境电商进口申报数据。考虑到模型验证的可操作性与专家评分的可行性，从中随机选取 100 条样本作为核心分析对象，这些样本覆盖管理风险、供应链风险、税收风险、产品风险及合规风险五个维度的 17 项具体指标。该数据集呈现典型的高维稀疏特征，大量字段以文本编码或类别标签形式记录，如海关监管代码、运输方式、产品税号等定性信息占据主导地位，这种非结构化的数据形态使得传统的数值评价方法面临适应性瓶颈。

针对上述数据特性，研究构建专家经验与机器学习协同的混合处理框架：首先依托海关领域专业知识对初始样本进行风险评分标注，建立高质量的标注数据集；继而利用机器学习方法对剩余样本进行自动化预测与批量处理，实现类别型特征向风险量化值的有效转换。这一策略既保证了数据标注的专业性与准确性，又充分发挥了机器学习在大规模数据处理中的效率优势。

3.2. 风险评价指标体系构建

跨境电商进口风险评价指标体系的构建需以海关监管理论与风险管理理论为基石，遵循系统性、代表性与可操作性的设计原则[9]。在准则层维度划定上，依据海关监管业务链条的关键节点特征，确立管理风险、供应链风险、税收风险、产品风险与合规风险五个核心评价维度。管理风险维度涵盖申报时间节点的时序分布规律与申报海关部门的地域监管属性；供应链风险维度包含途经高风险国家的轨迹标识、运输方式类别、不含包装重量的物理属性、配送服务提供商身份、申报主体身份特征、进口消费者身份要素；税收风险维度针对关税征管核心环节，设置项目税率适用准确性、税种分类合规性(如自由贸易协定优惠税率适用情况)及物品评估价值真实性三项关键指标；产品风险维度侧重商品归类与原产地合规性，具体包括原产国申报异常监测与六位数 HS 编码产品分类准确性验证；合规风险维度则指向贸易管制遵循程度，涉及原产地标识标注方式、关税支付类型区分(如即期付款与远期信用证)及进口用途代码(如国内消费原材料)要素。

Table 1. Cross-border e-commerce goods customs risk early warning indicator system
表 1. 跨境电商货物海关风险预警指标体系 [1]

准则层	指标层
管理风险	申报时间 C1
	申报海关部门 C2
供应链风险	途径高风险国家，已发运或计划发运的国家 C3
	运输方式 C4
	不含包装重量 C5

[1]C5、C10、C12 为定量指标，不作为专家评分目标，仅作为模型特征输入。

续表

	配送服务提供商(例如 DHL、FedEx)的风险 C6
	申报物品的人 C7
	进口该商品的消费者 C8
	向中国供应商品的海外商业合作伙伴 C9
税收风险	*项目税率* C10
	税种(例如，FTA 优惠税率) C11
	物品评估价值 C12
产品风险	原产国异常 C13
	6 位产品代码(例如 090121 = 咖啡，烘焙，未脱咖啡因) C14
合规风险	表明原产国的方式(例如，包装上的标记) C15
	区分关税支付类型(例如，即期付款远期信用证) C16
	进口用途代码(例如，国内消费的原材料) C17

表 1 实现了对跨境电商进口环节从申报主体、物流轨迹、税收征管到商品属性的全方位覆盖，通过定性描述与定量测度的有机结合，为后续风险量化建模与阈值判定提供了标准化的指标体系支撑。

3.3. 数据预处理方法

3.3.1. 专家评分的策略选择

在海关数据评价中，定性指标(如运输方式、产品代码、海关部门等)的量化是建模的关键前提。传统做法通常由专家对整条交易记录给出一个综合风险分，然后训练回归模型进行预测。该方法虽然简洁，但存在两个不足：一是综合分掩盖了不同风险维度的贡献来源，不利于精准监管；二是当模型预测结果异常时，难以进行归因分析。

为弥补上述不足，本研究采用一种探索性的替代路径：由专家对每条样本中的每个定性指标独立给出百分制风险分(例如 C1_risk = 56，C2_risk = 65，……，C17_risk = 78)，而非仅给一个总分。这一策略的理论优势在于：可以精细定位风险来源(例如某条记录仅“产品代码”指标风险极高，其余正常)；便于海关根据业务重点灵活组合各指标风险，避免不同指标风险相互抵消。但需要强调的是，该策略建立在“每个定性指标的取值本身蕴含独立风险信息”的强假设之上，这一假设对于部分指标(如申报时间、海关部门)并不天然成立。因此，本文将本方法定位为一次探索性的方法论尝试，其主要目的是验证“专家独立评分 + 机器学习预测”的技术可行性，并为后续更完善的指标体系研究提供实验基础。

需要特别说明的是，本研究的专家评分对象仅限于定性指标(C1~C4, C6~C9, C11, C13~C17)。对于 C5 (不含包装重量)、C10 (税率)、C12 (评估价值)三个定量指标，其本身已是数值型数据，海关系统中可直接获取其偏离程度(如与历史均值比较、与申报价格偏差等)，无需再由专家进行主观的百分制风险评分。在实际建模中，这三个定量指标将作为特征输入而非预测目标，用于辅助预测其他定性指标的风险分。这样处理既避免了“对数值打分”的业务不合理性，也使得模型更加聚焦于定性信息的风险量化。

本文所构建的监督学习标签来源于专家评分，其本质反映的是专家对风险的主观判断而非客观风险本身。因此，模型学习的目标可以理解为“专家决策函数”的近似，而非真实风险函数。这一定义决定了本文方法更适用于辅助决策与经验复现场景，而非替代真实风险判定。

3.3.2. 数据集划分与专家评分实施

本研究从 954 条跨境电商进口申报数据中随机选取 100 条样本作为分析对象。按照前 60 条、后 40

条进行划分：前 60 条用于专家标注和模型训练，后 40 条用于模型预测效果检验(后 40 条各个因素的真实风险分同样由专家给出，但模型在训练阶段不可见)。邀请 3 位海关领域专家，依据表 1 所示的 14 个定性指标(C1~C4, C6~C9, C11, C13~C17)，对前 60 条样本的每个指标独立进行百分制风险评分(0 表示无风险，100 表示极高风险)。评分过程中，专家仅依据该指标在当前样本中的取值进行判断，不考虑与其他指标的交互。对于明显不具备独立风险含义的指标(如 C1 申报时间)，专家统一按照“时间异常程度”打分(例如节假日申报赋予较高分值)。最终取 3 位专家评分的均值作为该指标的标准风险分。图 1 展示了部分原始样本的格式，图 2 展示了专家评分后的数据片段。

C1	C2	C3	C4	C6	C7	C8	C9	C11	C13	C14	C15	C16	C17
2020/1/2	40	CN	10		7OSA8BP	2P3V50X	06JLM7J	A	CN	620462	G	11	21
2020/1/2	40	CN	40	VZIZJJ	5RQOVK6	5VOMCFU	U0867HI	FCN2	CN	871200	E	43	21
2020/1/2	40	CN	10		LP6EWRR	E808G5S		C	CN	940180	B	11	21
2020/1/2	40	CN	40	MWIDNS	GKIYMO3	L470GO7	MU7NSIW	A	CN	611030	Y	14	21
2020/1/3	40	US	40	W6UCD9	PF1L1IG	PT98SSF	9SZTR8D	CIT	CN	851770	Y	11	21
2020/1/3	40	CN	10	W6UCD9	XCIAAPR	USZKXON	CZXGQEJ	A	CN	670210	G	14	21
2020/1/6	40	CN	10		Q9ZG6R5	NKE45NQ	UMOA3XS	C	CN	851762	G	11	21
2020/1/6	20	US	10		9GVIG2T	1GZBEVE	BUSF98I	A	US	420299	E	11	21
2020/1/6	33	VN	40		M65J4QO	IG4I3ZW	4K4PV9F	FVN1	VN	610910	S	14	21
2020/1/6	20	CN	10		9GVIG2T	TOOEFIB	41W8AHR	CIT	CN	850440	E	43	21
2020/1/6	16	JP	40		F463TSU	J9ROZJL		C	JP	392350	S	43	21
2020/1/8	20	CN	40	5V1OQR	JSORSCD	J6WCSC3	D6V32XG	A	CN	071080	E	11	21
2020/1/9	40	CN	40		H3RWZGO	YLLBMD1	YBRGVO2	A	CN	702000	E	43	21
2020/1/9	10	CN	10		RZXCX7U	KGR47BK	Z46YLS4	A	CN	701399	E	14	27
2020/1/10	40	VN	40	MWIDNS	8R629JR	RKDII26	KIMCL10	FVN1	VN	620343	E	14	21
2020/1/10	16	JP	40		LX5UCF2	2KO455D	13FN8OP	C	JP	294190	N	14	21
2020/1/13	16	LT	30		FOVFD9M	J4BY6D2	H9OUDBA	FEU1	LT	621149	B	43	21
2020/1/13	20	MY	10		SCO3ZOS	OUDRKX8	PDIBHHF	FAS1	MY	210690	E	11	21
2020/1/13	151	CN	10		DDNN5F9	HOKO6Q3	TNBC1SB	E1	CN	620640	G	11	28
2020/1/13	40	CN	40	VZIZJJ	9Z8AYH4	96FHRL1	7ZJ78BT	E1	CN	620463	G	43	21
2020/1/14	20	AT	10		7GHJ33J	L8LD7BY	K1KGVXC	CIT	HU	901510	B	11	21
2020/1/14	30	DE	10		SWTJELT	3B0TONJ	QT180W8	FEU1	DE	842191	E	11	21
2020/1/15	40	RU	10		HLQTVD1	708A72F	Q1RHL6S	A	RU	270111	Y	43	21
2020/1/15	20	CN	40		RQQ6ONT	L8LD7BY	CXPPJ83	FCN1	CN	950629	Y	11	21
2020/1/15	80	JP	10	3OGFAP	M96WZP8	YZQRS9F	Q7ML2OL	A	JP	848210	B	43	21
2020/1/15	40	DE	40	MWIDNS	SPKIA9H	SRJKKXK	KDT3IEI	A	PL	851539	E	14	21
2020/1/15	16	CN	40		43MSLZK	QN9P2J0	WXFKS7I	A	CN	071080	Y	18	21
2020/1/16	30	CN	50	MWIDNS	EFJR7PT	V981F2H	47SGJPL	FCN1	CN	842310	E	11	26
2020/1/17	80	CN	10	A5IT1U	X6Z3FIF	TC78TU0	5YE5KJM	A	CN	620640	G	14	21

Figure 1. Format of some original samples
图 1. 部分原始样本的格式

C1	C2	C3	C4	C6	C7	C8	C9	C11	C13	C14	C15	C16	C17
61	60	62	80	71	40	42	53	71	63	62	80	60	60
66	63	69	68	76	77	86	88	100	66	78	62	79	67
62	62	65	82	71	44	43	68	73	65	64	90	82	66
61	62	62	61	48	70	55	50	71	64	68	52	62	62
82	62	69	66	68	68	87	89	85	64	87	55	69	68
81	69	68	85	67	96	84	88	79	66	73	87	65	62
70	61	63	83	71	66	51	52	73	64	66	81	63	65
70	74	60	81	72	65	53	54	70	51	60	60	65	62
78	88	75	69	79	95	89	85	100	66	94	75	66	61
78	77	66	88	77	74	85	89	88	65	97	68	89	69
79	86	75	69	76	77	89	69	76	69	87	79	88	64
50	72	62	61	43	43	51	52	71	60	61	60	65	60
52	64	63	66	70	60	51	55	70	61	64	65	72	61
66	69	64	76	68	73	60	64	76	61	67	73	68	62

Figure 2. Data snippets after expert scoring
图 2. 专家评分后的数据片段

为确保专家评分的一致性与可靠性，本研究在正式评分前制定了详细的评分指南和量表。该指南针对 14 项定性指标分别明确了风险判定的核心维度与分值锚定标准：以 0 分表示“无风险/常规状态”，65 分表示“中等关注/需进一步核实”，100 分表示“极高风险/明显异常”。在实际操作过程中对 C3 (途经

高风险国家)以“是否途经国际制裁名单国家或高风险地区”为判定依据，对 C7 (申报物品的人)则综合考量“申报主体历史信用记录、企业规模与申报频次匹配度”因素。评分过程中，3 位专家独立打分，互不干扰。

评分完成后，从一致性和变异性两个维度评估评分者信度。首先，计算 3 位专家两两之间的 Pearson 相关系数(用来衡量线性相关程度)，结果显示专家 A 与 B、A 与 C、B 与 C 的相关系数分别为 0.84、0.81、0.86 ($p < 0.001$)，均高于 0.80 的可接受阈值，表明专家间具有较高的一致性；其次，计算各指标 3 位专家评分的标准差，C3 (途经高风险国家)、C13 (原产国异常)、C15 (原产国标识方式)等指标平均标准差为 5.3 分，变异较小，而 C7 (申报物品的人)、C8 (进口该商品的消费者)平均标准差为 9.5 分，变异相对较大，这与主体身份类指标本身的情境敏感性相符。最后综合考虑相关系数与评分变异程度，14 项指标的评分数据整体可靠，可作为训练标签用于后续建模。对于变异较大的 C7、C8 指标，已通过专家复核确认分歧源于业务场景复杂性而非标准理解偏差，故保留于指标体系中。

3.3.3. 特征工程与降维处理

由于样本量较小(60 条训练)，直接对所有定性特征进行独热编码会导致特征维数远超样本量，极易过拟合。为此，采取以下降维措施：

首先是日期特征提取，将 C1 (申报时间)转换为年、月、星期三个数值特征，删除原始日期列；接着是高基数类别合并，对 C14 (6 位 HS 编码)，仅保留前 2 位(HS 章节编码)，将类别数从数百降至约 20。对 C2 (海关部门)、C3 (途经国家)等取值超过 10 种的类别特征，采用频数编码(以该取值的出现频次代替原始类别)；最后是进行低基数类别独热编码，对 C4 (运输方式)、C6 (物流商风险)、C11 (税种)、C15~C17 等取值种类 ≤ 10 的特征，使用独热编码上述预处理通过 Column Transformer (用于对数据的不同列应用不同预处理流程的核心工具，是实现复杂数据预处理管道的关键组件)集成至 Pipeline (用于串联多个处理步骤的核心工具，确保数据预处理和模型训练以正确顺序执行，并防止数据泄露)中，确保数据不泄露。

3.4. 模型构建与实现

3.4.1. 建模任务定义

本研究的目标是：基于 60 条已标注样本，训练一个多输出回归模型，能够对新的申报记录自动预测其 14 个定性指标各自的风险分。模型输入为预处理后的特征向量(维度约 50~80)，输出为 14 维向量，对应 C1、C2、C3、C4、C6、C7、C8、C9、C11、C13、C14、C15、C16、C17 的风险分。

3.4.2. 模型选择与正则化

考虑到训练样本有限(60 条)，选择岭回归(Ridge)作为基回归器[10]，利用 L2 正则化抑制过拟合。将岭回归包装在元估计器(MultiOutputRegressor)中，以支持多输出。正则化系数 $\alpha = 1$。

3.4.3. 训练与预测流程

基于 Python 的 scikit-learn 库实现以下流程：

首先是加载数据，从 10,000 条样本(100fulldata.xlsx)中读取前 60 条样本的特征 X_train (60 行 × 14 列)，从 60 条专家评分后样本(60all.xlsx)中读取对应的 14 维风险分矩阵 y_train (60 行 × 14 列)；然后构建预处理管道，按照 4.3.3 节定义 ColumnTransformer，完成特征提取、降维、编码和标准化；接着构建完整管道并训练模型，通过 pipeline.fit (X_train, y_train)实现；最后是预测后 40 条数据，对 100fulldata.xlsx 中的后 40 条样本执行相同的特征工程，调用 pipeline.predict ()得到预测的 14 维风险分矩阵，保存为 predictions_40.xlsx (图 3 预测后数据片段)。

C1	C2	C3	C4	C6	C7	C8	C9	C11	C13	C14	C15	C16	C17
64.60279	66.94027	64.94356	69.78794	66.25814	67.3563	65.29935	66.74139	80.22441	61.77646	70.23621	63.45578	69.85854	63.5228
75.69172	71.02264	65.71474	80.20923	74.94964	70.32977	65.44635	67.2365	77.31849	60.08135	71.3891	73.43999	67.3262	63.64028
70.26882	71.94046	66.4497	74.48656	73.42582	67.96076	66.59541	67.91028	80.08603	61.17703	73.86276	69.21053	71.58126	63.77555
68.80226	66.28356	67.02539	74.19581	70.39791	69.20765	66.56045	68.68845	79.35369	66.08484	75.07429	75.01879	72.34595	64.41415
68.82613	68.23634	67.67329	69.61886	70.59165	71.21119	72.25572	71.97658	81.64413	65.24293	78.81808	66.63241	74.99718	64.659
72.17124	69.24168	67.64256	71.07699	70.0981	68.51345	72.89447	72.30501	79.70472	62.8595	74.55956	67.67921	72.39056	64.46106
67.54612	68.23617	68.45718	70.20474	72.27599	69.90887	71.97932	73.33489	82.55522	66.26232	78.17976	71.33499	75.41542	64.13639
65.57346	65.95089	66.3101	71.73595	69.4301	60.79043	60.88107	68.23811	79.0132	63.78837	71.70262	71.78599	72.32235	64.0643
70.79893	69.42444	66.81154	69.93494	66.61272	73.71698	72.80221	70.51244	81.82602	63.8946	78.05935	63.34329	71.32863	64.55588
74.71105	71.82006	66.18795	77.9245	70.33812	74.65937	67.73275	65.42721	75.33936	61.59244	70.4609	74.20732	66.84591	62.67277
72.81567	68.31267	66.26551	76.27154	68.73418	73.93983	70.22859	72.02795	80.69027	64.14096	76.38838	70.5765	71.43199	64.86625
68.38796	69.78348	66.52157	71.24161	72.32533	66.69696	66.52045	67.30563	80.72472	61.91408	74.3805	66.34522	71.34642	63.97821
67.72747	70.19755	66.56295	72.78914	70.04911	73.24972	69.01339	66.69221	78.25018	64.26919	74.90631	70.5486	71.80644	62.94436
71.35442	66.62386	67.51584	70.75431	69.35599	74.79284	77.27057	78.36801	84.34707	64.59109	78.83687	65.49963	72.31991	65.19412
68.07135	73.0001	67.9302	67.45203	67.75565	70.99377	74.17007	67.45858	77.13453	64.2279	76.1853	65.01431	77.21843	63.24765
64.01001	64.72045	64.57306	69.20421	62.51349	63.90583	63.79236	67.42843	77.23648	62.26288	68.80909	64.89474	67.75768	63.11939
71.18964	75.72501	67.95965	71.40128	70.83657	75.87376	76.86986	69.56824	78.61271	62.75363	76.34733	67.27919	75.19091	62.90363
71.63454	72.92004	68.19021	71.82865	70.48051	76.28239	75.77803	72.52909	80.65591	64.06543	77.48244	69.50794	72.82297	63.16311
70.27199	71.70637	67.28758	70.32326	72.36912	67.34374	70.00637	69.48612	81.40949	61.62267	76.37033	65.04337	71.52731	64.00462
73.1886	72.62896	66.09681	75.68081	70.07046	68.32743	64.8804	62.44168	72.99676	61.58083	72.29518	71.21951	70.20859	63.00573
70.30506	63.67499	67.30581	70.76154	65.66249	63.50439	68.28239	72.82954	78.46721	66.12527	75.37306	70.69578	73.7231	65.68624
72.29358	66.66924	65.98228	77.57444	71.9659	70.31503	65.35518	69.23767	77.82313	62.56711	70.58151	75.00584	67.35499	63.6573
67.10692	65.42498	63.54944	76.36416	67.00193	62.53145	55.84981	61.87797	73.29397	61.61473	67.65258	71.1744	67.81677	63.26393
65.51921	68.58234	65.46157	72.38751	68.28121	70.6853	68.22532	69.27179	79.10524	62.81332	72.09231	67.26065	72.9773	63.26072
71.07681	63.26247	65.19902	79.24065	70.59342	66.61191	68.35544	78.52109	81.32062	63.70509	74.15745	72.79541	72.43757	65.54698

Figure 3. Post-prediction data fragments
图 3. 预测后数据片段

3.4.4. 探索性尝试的说明

上述建模过程基于“每个指标可独立预测”的探索性假设。由于该假设在业务上并非普遍成立，且训练样本量偏小，本研究的模型预测结果应被视为初步的概念验证，而非可直接部署的成熟工具。其核心价值在于：验证了“专家独立评分 + 多输出回归”在海关数据场景中的技术可行性；暴露了指标独立假设与现实风险耦合之间的矛盾，为后续研究指明改进方向(例如引入交互特征或改用综合风险评分)。

此外，由于训练样本规模较小(60 条)且数据来源包含合成数据，模型在测试集上表现出的较高精度具有一定理想化特征。该结果更多反映模型对当前数据分布及专家评分规则的拟合能力，而非对真实复杂业务环境的泛化能力。因此，相关实验结果应主要作为方法可行性的验证，而非性能上限的证明。

4. 结果分析与讨论

本研究基于 60 条专家标注样本，训练了一个多输出岭回归模型，用于预测 40 条测试样本中 14 个定性指标(C1~C4, C6~C9, C11, C13~C17)各自的风险分。以下从数据处理效率、预测精度、方法局限性三个维度进行分析。

4.1. 数据处理效率

在数据处理效率方面，模型训练(含特征工程与交叉验证选参)耗时约 2.3 秒，完成对 40 条测试样本的批量预测耗时约 0.1 秒。若完全依赖人工评分，40 条样本需 3 位专家每人约 20 分钟。相比之下，机器学习方法将评分时间从小时级压缩至秒级，显著提升了处理效率。这一优势在海关通关现场具有实际意义——当面临大量申报数据时，自动化预测能够大幅减轻人工审核负担。

4.2. 预测精度

为评估预测精度，本研究采用平均绝对误差(MAE)作为评价指标分别计算了模型在 40 条测试集上对 14 个指标各自预测的平均绝对误差(MAE)以及整体平均 MAE。同时，以“预测训练集各指标均值”作为简单基线进行对比，MAE 的计算公式如(1)所示：

$$\mathrm{MAE}=\frac{1}{n}\sum_{i=1}^{n}\left|y_i-\hat{y}_i\right| \tag{1}$$

模型在测试集上的整体平均 MAE 为 0.97 分。各指标 MAE 如表 2 所示。

Table 2. Mean absolute error (MAE) of predictions for each qualitative indicator
表 2. 各定性指标预测平均绝对误差(MAE)[2]

指标	MAE
C1	0.98
C2	0.94
C3	0.13
C4	0.89
C6	1.11
C7	1.88
C8	1.53
C9	1.63
C11	0.69
C13	0.31
C14	1.77
C15	0.31
C16	1.08
C17	0.38

由表 2 可以看出，模型能够较高精度地学习专家对定性信息的风险判别逻辑。与简单均值基线相比，模型误差显著降低，验证了多输出回归在定性指标风险量化中的有效性和可行性。

由表 2 还可以看出，C7 (申报物品的人，MAE = 1.88)、C14 (6 位产品代码，MAE = 1.77)、C9 (海外商业合作伙伴，MAE = 1.63)及 C8 (进口该商品的消费者，MAE = 1.53)预测误差较大。这些指标均涉及主体身份属性或基数比较大的类别编码，风险判别比较依赖历史行为数据与情境化经验，专家评分变异较大，模型学习难度相应增加。相比之下，C3 (途经高风险国家，MAE = 0.13)、C13 (原产国异常，MAE = 0.31)等指标准确率高，其风险判定规则明确(如是否途经制裁国家)，专家共识度高。

进一步分析测试集中预测误差最大的样本发现，高误差多集中于多风险因素场景(如“节假日申报 + 高风险原产国 + 特殊用途代码”)。独立评分假设下，模型难以捕捉此类交互效应，导致复合风险场景预测偏差。未来可通过引入交互特征或转向综合风险评分策略加以改进。

需要说明的是，尽管模型在测试集上取得了较低的平均绝对误差，但该结果可能被以下因素影响：数据样本规模较小，训练集与测试集分布差异有限；部分特征经过编码处理后具有较强结构性，使模型较易捕捉模式；标签来源于专家评分，存在一定规律性。因此，该精度结果应谨慎解读，不能简单等同于模型在真实业务环境中的表现。

4.3. 方法局限性与探索性定位讨论

上述结果验证了“专家独立评分 + 多输出回归”在海关数据场景中的技术可行性，且预测精度达到

[2]单位为(得)分。

了较高水平，但仍需指出该方法有着以下的局限性：

指标独立假设的简化性，本方法建立在“每个定性指标的取值本身蕴含独立风险信息”的假设之上。然而，实际风险往往是多指标耦合的结果(例如“深夜报关 + 高风险原产国”)。独立评分模型无法显式捕捉这种交互效应，但在本实验中由于指标间相关性较弱，独立假设并未造成明显误差。

样本量偏小，60 条训练样本对于多输出回归(输出维度 14)而言基本够用，但经过独热编码后特征维度达 320，样本量仅为特征维度的 1/5，存在一定的过拟合风险。

线性模型的表达能力可能不足，当前采用线性岭回归，若指标间存在强非线性关系，模型可能表现不足，未来可尝试随机森林或梯度提升树，但需配合严格的降维和正则化。

专家评分一致性的潜在影响，本研究采用 3 位专家评分的均值作为标签，但未对专家间一致性进行量化分析。若专家对某些指标(如 C7 申报物品的人、C8 消费者)分歧较大，会降低监督信号的质量。后续可通过计算组内相关系数(ICC)评估一致性，并考虑使用加权平均或专家置信度。

4.4. 探索性尝试的价值总结

尽管存在上述局限，但本研究成功搭建了从原始定性数据到 14 个定性指标风险预测的端到端 Pipeline，证明了“专家知识标注–特征工程–多输出回归”在海关数据场景中的可实施性，且预测精度达到较高水平；同时使方法得到有效性量化，通过严格的训练/测试集划分和平均绝对误差评估，定量证明了多输出岭回归能够有效学习专家对定性指标的风险评分逻辑，为海关风险评价的自动化提供了可行方案；更重要的是所构建的 60 条标注数据集和预处理代码可作为后续更复杂模型(如非线性集成学习)的对比基准。未来工作可以聚焦于以下四个方面：引入指标交互特征；采集更大规模的标注数据；对比非线性模型与集成方法的性能；探索将预测结果用于综合风险评分(如加权融合)的实际应用方案。

5. 总结

本文从机器学习与统计学交叉视角，系统探讨了机器学习在海关数据评价中的应用，构建了“专家评分–机器学习预测”的混合模型，实现了定性指标的定量化处理，验证了其在风险量化与专家经验模拟中的可行性。研究表明，机器学习能有效提升海关数据评价的自动化水平与风险识别能力。尽管存在数据规模、模型复杂度等不足，但本文为“经验知识的数据化表达”提供了初步参考。未来将引入多模型对比、专家一致性度量及更大规模真实业务数据，以提升方法的稳健性与应用价值。

基金项目

2024 年国家级大学生创新创业训练计划项目(项目编号：202410274010)受上海海关学院大学生创新创业训练计划项目经费支持。

参考文献

[1] 闫宇宁, 苏晓伟, 万振龙. 深度学习技术在海关风险甄别中的应用研究[J]. 中国口岸科学技术, 2020(3): 45-51.

[2] 人工智能和机器学习应用于海关工作中的研究报告[R]. WCO “智慧海关”合作项目, 2025. https://scp.wcoomd.org/reports, 2026-3-30.

[3] 侯艳艳. 量子机器学习的回归和分类算法研究[D]: [博士学位论文]. 北京: 北京邮电大学, 2023.

[4] 王光滔, 赵雯, 江宇静, 等. 机器学习在地表水水质管理中的应用[J]. 环境工程学报, 2024, 18(11): 3035-3048.

[5] 侯敏. 量子机器学习算法研究概述[J]. 通讯世界, 2024, 31(8): 139-141.

[6] Jeong, C., Kim, S., Park, J., *et al*. (2022) Customs Import Declaration Datasets. https://arxiv.org/abs/2208.02484

[7] Xu, L., Skoularidou, M., Cuesta-Infante, A., *et al*. (2019) Modeling Tabular Data Using Conditional GAN. *Annual*

Conference on Neural Information Processing Systems 2019, *NeurIPS* 2019, Vancouver, 8-14 December 2019, 7335-7345.

[8] 虞苏妍, 左芳芳, 田盼. 人工智能的创新基石: 合成数据[J]. 中国工业和信息化, 2024(11): 10-14.

[9] 王翔, 李志鹏, 刘萤, 等. 应用世界海关组织数据模型构建智能边境信息框架探究[J]. 中国口岸科学技术, 2023(4): 15-22.

[10] Hoerl, A.E. and Kennard, R.W. (1970) Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, **12**, 55-67. https://doi.org/10.1080/00401706.1970.10488634