2025 年 MathorCup 大数据竞赛 B 题赛题解析

一、赛题整体背景与核心任务拆解

1.1 赛题本质

赛道 B 聚焦**物流售后理赔的风险管控与成本控制**，核心是通过历史运单数据建立规则与模型，**解决“风险标注”“赔付金额预测”“风险类型预测”三大任务**，最终为物流企业提供“快速识别不合理诉求、精准预测赔付成本”的决策支持。其核心矛盾在于平衡“客户体验”与“企业成本”，明确的业务逻辑展开。

1.2 核心数据与字段意义

数据附件	内容说明	核心用途
附件 1	已完成谈赔的历史运单，含实际赔付金额(P)、索赔金额（C）、商品属性（类型/保价）、运输属性（超时时长/异常原因）、网点属性（发单量/理赔率）等	规则建立、模型训练（问题 1-3 的基础数据）
附件 2	待预测运单，缺失实际赔付金额(P)，特征字段与附件 1 一致	结果预测（问题 2 需预测 P，问题 3 需预测风险类型）

二、数据预处理

2.1 数据列名规整与类型定义

A-C 列、I 列、K-M 列、O-R 列(类别型)：线路类型，是否 c2c，生鲜妥投，寄件是否内部，异常原因，进线渠道，妥投到进线时长，商品类型，新旧程度，寄件 B/C，进线人身份。

D 列、J 列、N 列、Y 列(数值型)：保价金额，配送超时时长，索赔金额，实际赔付金额。

E-H 列(高基数 ID 类)：始发城市 ID，目的城市 ID，寄件人 id，收件人 id。

S-X 列(网点统计数值型)：始发网点发单量，始发网点万单理赔率，始发网点赔付比例，目的网点发单量，目的网点万单理赔率，目的网点赔付比例。

2.2 缺失值处理

异常值可能对模型有不良影响，可做处理包括：

配送超时时长(J 列)：缺失值很可能代表“准时送达”，可填充为 0。

异常原因(K 列)：缺失值代表“无提报异常”，可以创建一个新类别，如“NoAbnormality”，这本身就是一个有用的信息。

网点特征(S-X 列)：如果有缺失，可能是新网点无历史数据，可使用全局的中位数填充，因为它对异常值不敏感。

2.3 异常值检测与平滑

重点检查：保价金额(D)、索赔金额(N)和实际赔付金额(Y)。

方法:使用箱线图（BoxPlot）识别离群点。对于那些远超正常范围的值（例如，索赔金额远大于保价金额数倍且商品并非贵重品），可以采用盖帽法，例如将所有超过 99.9%分位数的值替换为 99.9%分位数的值，以减小极端个例对模型的负面影响。

2.4 数据标准化

标准化有助于消除不同特征之间的量纲差异，使其对模型影响均衡。

三、问题 1：基于附件 1 的风险标注模型建立（规则制定任务）

3.1 核心目标

通过“索赔差额（D）”和“实际赔付金额（P）”，将附件 1 运单划分为**“合理诉求”“诉求偏高”“严重超额”**三类，需同时满足：

（1）硬约束：合理诉求占比≥85%，严重超额占比<3%；

（2）业务逻辑：P 越高，D 的合理阈值越宽松；同类运单 D 分布需“密集”（方差小），不同类 D 密集度差异较显著。

3.2 关键解题步骤（无监督规则制定，非机器学习）

3.2.1 数据预处理：先算 D+清异常

（1）第一步：计算索赔差额 D

严格按初赛文件公式：索赔差额（D）=实际赔付金额（P）-索赔金额（C）。

关键解读：D 通常为负（客户索赔金额 C 普遍高于企业实际赔付 P），D 越接近 0（或为正），代表客户诉求与企业赔付标准越契合，合理性越强；D 负得越多，诉求越不合理。

（2）第二步：清洗异常值与缺失值（易忽略异常值对规则的扭曲）

3.2.2 探索性分析（EDA）：验证业务规律（易跳过 EDA 直接定规则，导致规则脱离数据）

必须通过 EDA 揭示 P 与 D 的关联，为规则提供数据支撑，核心分析维度及细节如下表：

分析内容	操作方式	意义与预期结论
P-D 散点图	（1） 提取附件 1 中 P 和 D 列； （2）横轴设为 P，纵轴设为 D； （3）可按“商品类型”着色，例：Jewelry→红、Miscellaneous Goods→蓝； （4）其他醒目颜色标注典型异常点。	验证“P 越高，D 阈值越松” （1）相同颜色的点在 P 轴上应该多数集中在某个区间（同一商品类型的实际赔付金额集中） （2）随 P 增大，D 的“合理分布区域”拓宽
D 的分层分布	（1）按 P 分档（例：0~120、120~350、350~800、800~2000、	验证“同类 D 密集，不同类差异显著”

	2000+）； （2）每档绘制 D 的核密度图； （3）标注“合理/超额”预期区间	（1）合理诉求区域：核密度图峰度高，即密度高； （2）严重超额区域：峰度低，即密度低
比例统计	（1）计算全量 D 的 3%、85%分位数； （2）统计各分位数区间样本占比； （3）对比硬约束（合理≥85%、超额 <3%）	初步判断阈值范围，避免方向偏差

3.2.3 规则制定（即风险标注模型）

核心逻辑：**按 P 分层→每层内按 D 分位数定阈值**（易一刀切定阈值，忽略 P 对 D 的影响），具体步骤如下：

（1）**分层：**分层目标是确保每层 P 范围连续、样本量均匀，避免阈值失真，可选用 **K-means 聚类和分位数分层**。

（2）**每层 D 阈值确定**

分层方法	优点	缺点	适用场景
K-means 聚类	自动识别 P 自然分布，客观性强	对初始中心敏感，需多次运行；P 分布不均时效果差	数据量大
分位数分层	操作简单，每层样本量均匀，无需调参	主观性强；可能割裂 P 自然分布	数据量中/小

结合方案参考：

方案 1：DBSCAN 聚类划分 P 场景+分位数回归法定动态阈值

方案 2：GMM 聚类划分风险粗类别+分层分位数法细化阈值

（3）规则验证（易不验证直接应用，导致规则不符合业务）

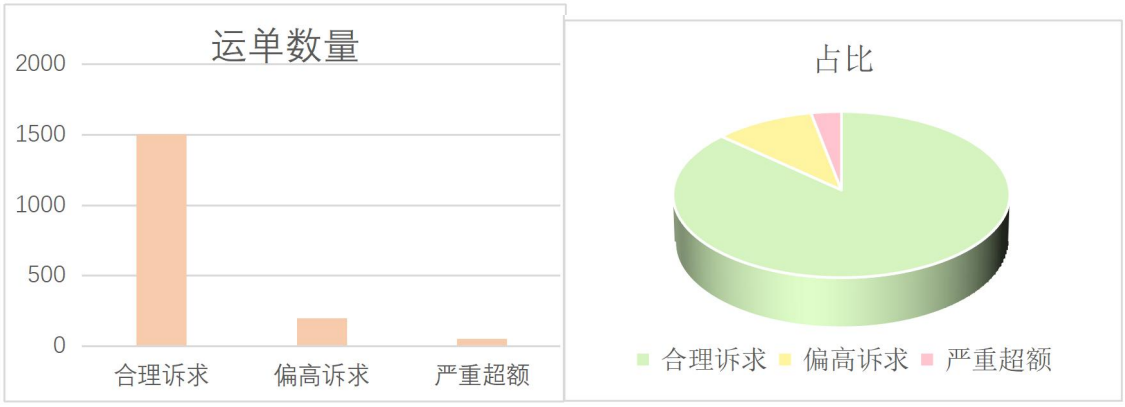
比例验证：全量样本三类占比需满足“合理≥85%、超额＜3%”，不满足则微调分位数；

密集度验证：计算同类 D 方差，需满足“合理方差＜偏高方差＜超额方差”。

3.2.4 输出附件 1 标注结果

生成“运单号-PCD-风险类型-所属 P 档”对照表，“风险类型”严格填写“合理诉求”“诉求偏高”“严重超额”，为后续问题 3 提供训练标签。

在论文中可进行图像可视化（可用柱状图、散点图等）和表格清晰化展示，如下为示意图和表：



风险标注	运单数量	占比
合理诉求		
偏高诉求		
严重超额		

四、问题 2：附件 2 实际赔付金额预测（回归模型任务）

4.1 核心目标

用附件 1 的运单特征建立回归模型，预测附件 2 的实际赔付金额，兼顾“预测精度”与“业务适配性”。

4.2 关键解题步骤

4.2.1 特征工程（易特征泄露或冗余，导致模型泛化差）

特征筛选：不参考无价值特征（运单号、寄件人 ID），保留核心特征（商品类型、保价金额、异常原因等）；删除冗余特征。

特征处理（按类型针对性处理）

特征类型	示例特征	处理方法
分类特征（低基数）	商品类型（Fresh/Electronic）、异常原因（Damage/Lost）	独热编码（One-Hot）
分类特征（高基数）	始发城市 ID、目的城市 ID（类别>50）	目标编码：用该类别下 P 的均值替代类别
数值特征	保价金额、配送超时时长	异常值：IQR 法盖帽；标准化
二值特征	是否为 C2C（0/1）、是否生鲜妥投及时（0/1）	直接保留 0/1 编码

衍生特征（提升模型解释力）

基于业务场景构建新特征，例如：保价金额/索赔金额——反映客户保价与索赔的匹配度；始发网点理赔率+目的网点理赔率——反映链路风险；配送超时时长/预计配送时长——标准化超时程度。

4.2.2 模型选择训练与预测评估

（易盲目选复杂模型，忽略可解释性与效率）

（1）回归任务选择

需可捕捉非线性、对大误差敏感、效率高的模型，可以使用模型包括：线性回归、随机森林回归、LightGBM 回归或 XGBoost。

（2）特征重要性分析

训练后，查看模型给出的特征重要性排序。我们预期索赔金额(N)以及我们构建的

寄件人_历史平均索赔差额等用户画像特征是否名列前茅。既能验证我们思路的正确性，也能写入论文作为亮点。

（3）训练与评估

采用 5 折或 10 折交叉验证进行训练和评估，以获得模型性能的稳健估计并进行调参。

评估指标：可使用 **RMSE(均方根误差)、MAE(平均绝对误差)、R^2（拟合优度/决定系数）。**

（4）预测

用全部处理好的附件 1 数据训练最终模型，然后对处理好的附件 2 数据进行预测。

五、问题 3：附件 2 风险标注预测与方法对比

5.1 核心目标

预测任务：用附件 1 的标注结果训练分类模型，直接预测附件 2 的风险类型（和问题 1 的三类标签一致），结果填入 Result 文件。

分析任务：处理“严重超额样本占比极低”的不平衡问题（如 SMOTE 过采样、调整类别权重），对比两种风险预测方法的优劣——“间接法”（用问题 2 的预测赔付金额+问题 1 的规则）和“直接法”（问题 3 分类模型直接预测风险）。

5.2 关键解题步骤

5.2.1 数据准备（易特征泄露，导致模型虚假高性能）

标签：附件 1 的风险类型（0=合理诉求，1=诉求偏高，2=严重超额），确保标签与运单特征一一对应，可通过运单号关联；

特征：与问题 2 完全一致；

数据划分：附件 1 按 8:2 划分为训练集（80%）和验证集（20%），采用分层抽样（按风险类型比例抽样），确保验证集与训练集的风险类型分布一致，避免验证集无严重超额样本。

5.2.2 处理“严重超额样本不均衡”

严重超额样本占比<3%，直接建模会导致模型偏向多数类，即合理诉求，漏判严重超额单（漏判会导致企业超额赔付，成本风险高），可从“数据层面”和“算法层面”双重处理：

处理层面	方法	优点	缺点
数据层面	SMOTE 过采样	（1）避免随机过采样导致的过拟合； （2）生成“真实”少数类样本，保留特征分布	（1）若少数类样本本身有噪声，会放大噪声； （2）对高维数据，插值样本可能与多数类重叠
数据层面	欠采样	操作简单，训练速度快	（1）可能丢失有效信息 （2）易导致模型欠拟合，泛化差
算法层面	类别权重调整	（1）无需修改数据，直接在训练中惩罚少数类误分类 （2）适配树模型，无额外计算成本	（1）若少数类样本量极少，权重过高可能导致过拟合； （2）对线性模型效果有限
混合策略（推荐）	SMOTE + 类别权重	双重保障，显著提升少数类召回率，且降低过拟合风险	操作稍复杂，需调试 SMOTEk 值

5.2.3 分类模型训练与预测

（1）模型选择

模型类型	优点	缺点
LightGBM 分类器	（1）处理非线性能力强； （2）支持类别权重和特征重要性； （3）效率高	可解释性弱于逻辑回归
逻辑回归（多分类）	（1）可解释性强； （2）训练快	（1）无法捕捉非线性； （2）对不均衡数据敏感
随机森林分类器	（1）抗过拟合； （2）对异常值不敏感	（1）高维数据效率低； （2）少数类召回率低于 LightGBM

（2）评估指标选择（不用准确率评估不均衡数据）

准确率（Accuracy）对不均衡数据无意义，需选择以下指标：宏 F1（Macro-F1）、**少数类召回率（Recall_超额）、加权 F1（WeighteD-F1）。**

5.2.4 方法对比：间接法 vs 直接法

需从“业务需求”“技术性能”“落地成本”三方面对比，结合业务逻辑给出建议：

路径一(回归+规则):

优势：白盒逻辑，强可解释性。便于业务落地和规则迭代。

劣势：误差累积，问题 2 回归模型的预测误差会直接影响最终分类的准确性。

路径二(直接分类):

优势：端到端优化，潜在性能更强，模型直接学习特征到风险等级的复杂映射，无需中间步骤。

劣势：黑盒模型，可解释性弱，且模型效果高度依赖问题 1 的标签质量。

六、竞赛技巧与论文撰写建议

先搭建基线：**先用一个基础模型快速跑通整个流程，得到一个初步结果**。然后再逐步迭代，更换更强的模型、尝试更复杂的增强策略。

可视化是王道：在论文中大量使用图片。包括：数据增强效果对比、模型结构图、特征图可视化，以及各种评估指标可视化。**一定要把结果做的准确又清楚!!!**

结果文件格式： 严格按照要求的格式输出。

七、MathorCup 大数据竞赛 B 题可参考相似赛题及优秀论文

2023 年 MathorCup 高校数学建模挑战赛—大数据竞赛 B 题

电商零售商家需求预测及库存优化问题（附件给了赛题和 4 篇获奖优秀论文）