**数据挖掘实验二 — 完整报告**

姓名：常欣然

学号：L12314126

日期：2025-10-27

# 原实验报告（合并自 experiment\_report\_final.docx）

数据探索（EDA）与类别分布分析

电商 类别分布

电商：发现与分析

发现：电商 存在严重的类别不平衡，类别“平板”占主导（10000 / 62774），而“热水器”仅有 575 条样本，比例约 17.4:1。

分析：这种不平衡将使模型倾向于多数类，导致少数类（如负面评论）被忽略，评价指标（如F1）对少数类不敏感。

酒店 类别分布

酒店：发现与分析

发现：酒店 各类别分布相对均衡（样本总数 7765）。

分析：可直接进行交叉验证与常规调参。

外卖 类别分布

外卖：发现与分析

发现：外卖 各类别分布相对均衡（样本总数 11987）。

分析：可直接进行交叉验证与常规调参。

数据不平衡处理与模型优化

电商 - SVM

解决方案：针对类别不平衡，采用了 随机过采样(文本复制) - 过采样(目标=多数类的50%)。

验证策略：5折交叉验证（StratifiedKFold）

电商 - NB

解决方案：针对类别不平衡，采用了 随机过采样(文本复制) - 过采样(目标=多数类的50%)。

验证策略：5折交叉验证（StratifiedKFold）

酒店 - SVM

解决方案：针对类别不平衡，采用了 所有少数类样本已达标，无需过采样。

验证策略：5折交叉验证（StratifiedKFold）

酒店 - NB

解决方案：针对类别不平衡，采用了 所有少数类样本已达标，无需过采样。

验证策略：5折交叉验证（StratifiedKFold）

外卖 - SVM

解决方案：针对类别不平衡，采用了 所有少数类样本已达标，无需过采样。

验证策略：5折交叉验证（StratifiedKFold）

外卖 - NB

解决方案：针对类别不平衡，采用了 所有少数类样本已达标，无需过采样。

验证策略：5折交叉验证（StratifiedKFold）

⚠️ 未找到 docs/新增代码与实现汇总.md