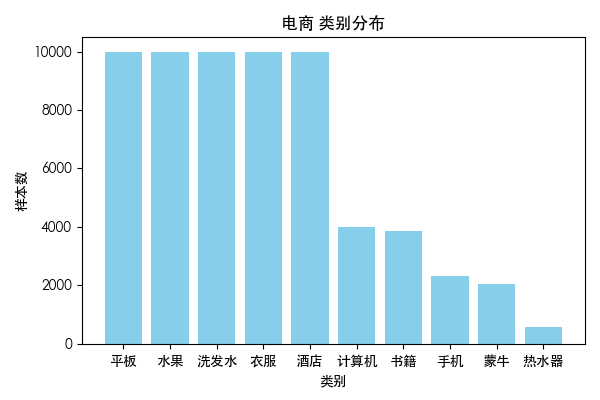
# 数据探索（EDA）与类别分布分析

## 电商 类别分布

|  |  |
| --- | --- |
| 类别 | 样本数 |
| 平板 | 10000 |
| 水果 | 10000 |
| 洗发水 | 10000 |
| 衣服 | 10000 |
| 酒店 | 10000 |
| 计算机 | 3992 |
| 书籍 | 3851 |
| 手机 | 2323 |
| 蒙牛 | 2033 |
| 热水器 | 575 |



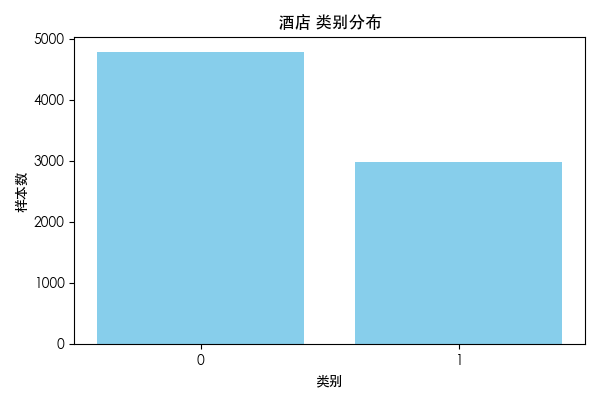
### 电商：发现与分析

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

|  |  |
| --- | --- |
| 类别 | 样本数 |
| 0 | 4789 |
| 1 | 2976 |



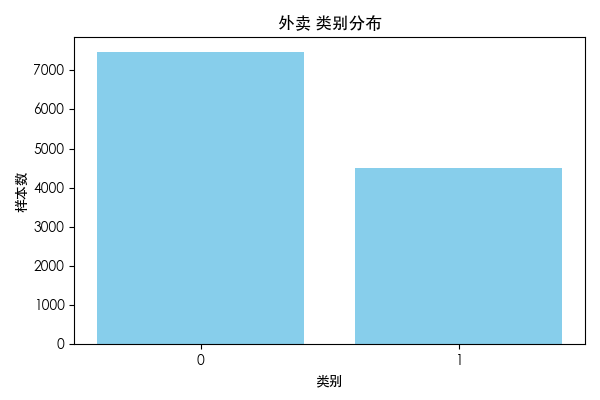
### 酒店：发现与分析

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

|  |  |
| --- | --- |
| 类别 | 样本数 |
| 0 | 7478 |
| 1 | 4509 |



### 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。

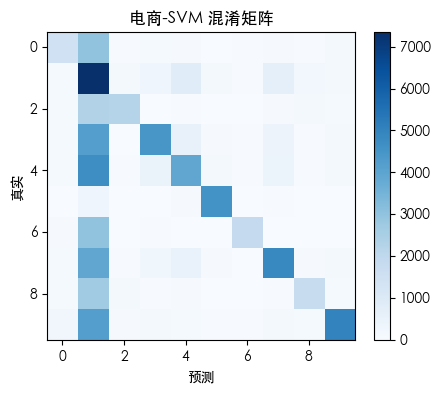
# 数据不平衡处理与模型优化

## 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）

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 类别 | precision | recall | f1-score | support |
| 书籍 | 0.628 | 0.298 | 0.404 | 5000.000 |
| 平板 | 0.207 | 0.736 | 0.323 | 10000.000 |
| 手机 | 0.821 | 0.438 | 0.572 | 5000.000 |
| 水果 | 0.766 | 0.445 | 0.563 | 10000.000 |
| 洗发水 | 0.634 | 0.391 | 0.484 | 10000.000 |
| 热水器 | 0.897 | 0.911 | 0.904 | 5000.000 |
| 蒙牛 | 0.877 | 0.372 | 0.522 | 5000.000 |
| 衣服 | 0.712 | 0.485 | 0.577 | 10000.000 |
| 计算机 | 0.722 | 0.355 | 0.476 | 5000.000 |
| 酒店 | 0.838 | 0.498 | 0.624 | 10000.000 |
| accuracy | 0.499 | 0.499 | 0.499 | 0.499 |
| macro avg | 0.710 | 0.493 | 0.545 | 75000.000 |
| weighted avg | 0.684 | 0.499 | 0.534 | 75000.000 |

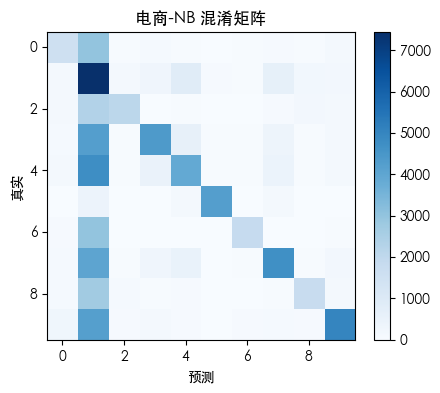


## 电商 - NB

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）

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 类别 | precision | recall | f1-score | support |
| 书籍 | 0.597 | 0.308 | 0.406 | 5000.000 |
| 平板 | 0.207 | 0.745 | 0.323 | 10000.000 |
| 手机 | 0.822 | 0.417 | 0.553 | 5000.000 |
| 水果 | 0.757 | 0.438 | 0.555 | 10000.000 |
| 洗发水 | 0.624 | 0.388 | 0.478 | 10000.000 |
| 热水器 | 0.953 | 0.847 | 0.897 | 5000.000 |
| 蒙牛 | 0.853 | 0.370 | 0.516 | 5000.000 |
| 衣服 | 0.714 | 0.471 | 0.568 | 10000.000 |
| 计算机 | 0.709 | 0.355 | 0.473 | 5000.000 |
| 酒店 | 0.819 | 0.499 | 0.620 | 10000.000 |
| accuracy | 0.492 | 0.492 | 0.492 | 0.492 |
| macro avg | 0.705 | 0.484 | 0.539 | 75000.000 |
| weighted avg | 0.678 | 0.492 | 0.529 | 75000.000 |

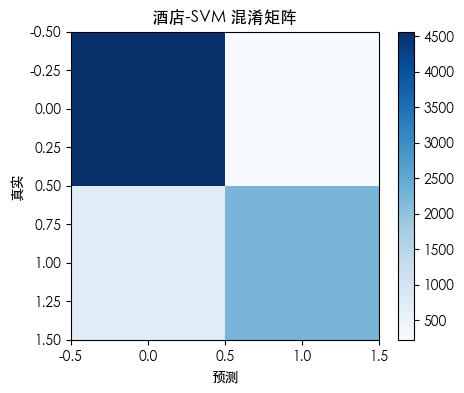


## 酒店 - SVM

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）

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 类别 | precision | recall | f1-score | support |
| 0 | 0.862 | 0.953 | 0.905 | 4789.000 |
| 1 | 0.908 | 0.754 | 0.824 | 2976.000 |
| accuracy | 0.876 | 0.876 | 0.876 | 0.876 |
| macro avg | 0.885 | 0.853 | 0.864 | 7765.000 |
| weighted avg | 0.880 | 0.876 | 0.874 | 7765.000 |

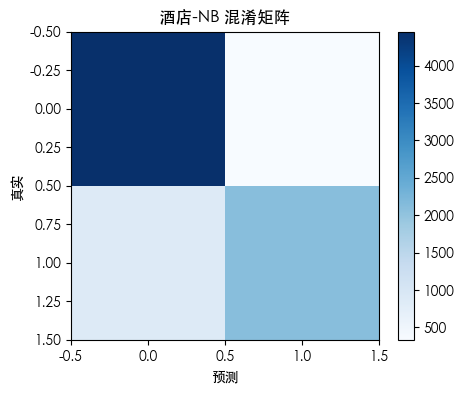


## 酒店 - NB

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）

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 类别 | precision | recall | f1-score | support |
| 0 | 0.836 | 0.930 | 0.881 | 4789.000 |
| 1 | 0.863 | 0.707 | 0.777 | 2976.000 |
| accuracy | 0.845 | 0.845 | 0.845 | 0.845 |
| macro avg | 0.849 | 0.819 | 0.829 | 7765.000 |
| weighted avg | 0.846 | 0.845 | 0.841 | 7765.000 |

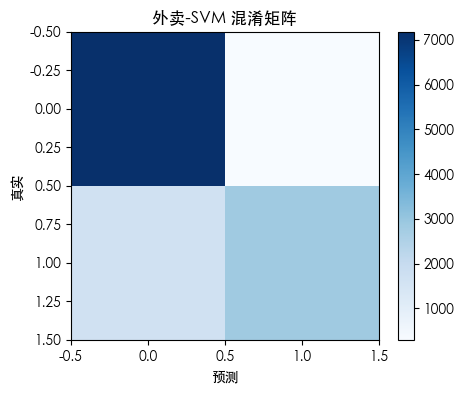


## 外卖 - SVM

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）

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 类别 | precision | recall | f1-score | support |
| 0 | 0.812 | 0.960 | 0.880 | 7478.000 |
| 1 | 0.905 | 0.632 | 0.744 | 4509.000 |
| accuracy | 0.837 | 0.837 | 0.837 | 0.837 |
| macro avg | 0.859 | 0.796 | 0.812 | 11987.000 |
| weighted avg | 0.847 | 0.837 | 0.829 | 11987.000 |



## 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）

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 类别 | precision | recall | f1-score | support |
| 0 | 0.791 | 0.961 | 0.868 | 7478.000 |
| 1 | 0.900 | 0.579 | 0.705 | 4509.000 |
| accuracy | 0.817 | 0.817 | 0.817 | 0.817 |
| macro avg | 0.846 | 0.770 | 0.786 | 11987.000 |
| weighted avg | 0.832 | 0.817 | 0.807 | 11987.000 |

