# 实验报告（草稿）

日期：2025-10-20

作者：自动生成（请校验与补充）

## 一、摘要

本实验对三个中文评论数据集（酒店 ChnSentiCorp\_htl\_all、商品评论 online\_shopping\_10\_cats、外卖评论 waimai\_10k）进行了统一预处理、自动情感打分（SnowNLP）、以及基于 TF-IDF 的两种基线分类器（MultinomialNB / LinearSVC）的训练与评估。重点工作：修复编码/转义问题、统一 UTF-8 输出、在训练集上做少数类上采样及分层拆分回退策略。本文档为实验结果草稿，含关键指标、代表样例、发现与下一步建议。

## 二、数据集简要

- hotel: 原始文件 `NLP数据集/酒店评论数据/ChnSentiCorp\_htl\_all.csv`（处理后 `data/processed\_hotel.csv`）

- ecommerce: 原始 `NLP数据集/电商评论数据/online\_shopping\_10\_cats.csv`（处理后 `data/processed\_ecommerce.csv`）

- waimai: 原始 `NLP数据集/外卖评论数据/waimai\_10k.csv`（处理后 `data/processed\_waimai.csv`）

输出文件（重要）：

- `data/processed\_{hotel,ecommerce,waimai}.csv`（UTF-8）

- `output/labels\_{hotel,ecommerce,waimai}.csv`（含 sentiment\_score, sentiment\_label）

- `output/classification\_report\_{\*\_nb|\_svm}.txt`

- `output/eda\_stats.csv`, `output/length\_hist\_{name}.png`

## 三、预处理与打标签

- 自动编码检测并回退：先检测（chardet），尝试编码顺序为[检测值, utf-8, gb18030, latin1]，并使用 `on\_bad\_lines='skip'` 提高鲁棒性；最终结果用 gb18030 成功读取原始文件（日志中有记录）。

- 文本清洗（`simple\_clean`）：剔除换行、尝试对包含文字型转义序列（如 `\u4e2d` / `\xE5`）进行 `unicode\_escape` 解码以恢复中文。

- 分词：使用 `jieba.cut()`，生成 `tokens` 列与 `tokens\_join`（以空格连接的分词结果，便于 TF-IDF）；写出 processed CSV 时强制 `encoding='utf-8'`。

- 自动标注：使用 SnowNLP 计算 `sentiment\_score`（[0,1]），阈值 0.5 -> `sentiment\_label`（1=正/非负，0=负）。

## 四、模型与训练流程（简述）

- 特征：TfidfVectorizer(max\_features=20000) 使用 `tokens\_join`。

- 模型：MultinomialNB；LinearSVC(class\_weight='balanced')（若训练单类则捕获异常并记录）。

- 不平衡处理：若训练集显著不平衡（少数类小于 max(5, maj\*0.2)），对训练集做上采样（resample）到多数类大小；分层拆分时在 `stratify` 不稳时尝试多次 random\_state 回退，并在必要情况下显式把一个少数类样本加入测试集以保证测试覆盖。

## 五、关键结果（classification\_report 摘要）

摘自 `output/classification\_report\_\*.txt`，下列为两类分数的主要指标（precision / recall / f1 / support）：

- hotel (NB)

- class 0: precision=0.7391 recall=0.9906 f1=0.8466 support=958

- class 1: precision=0.9665 recall=0.4370 f1=0.6019 support=595

- accuracy=0.7785

- hotel (SVM)

- class 0: precision=0.9135 recall=0.8706 f1=0.8915 support=958

- class 1: precision=0.8063 recall=0.8672 f1=0.8356 support=595

- accuracy=0.8693

- ecommerce (NB)

- class 0: precision=0.8666 recall=0.8999 f1=0.8829 support=6541

- class 1: precision=0.8863 recall=0.8494 f1=0.8675 support=6014

- accuracy=0.8757

- ecommerce (SVM)

- class 0: precision=0.9083 recall=0.9054 f1=0.9068 support=6541

- class 1: precision=0.8974 recall=0.9006 f1=0.8990 support=6014

- accuracy=0.9031

- waimai (NB)

- class 0: precision=0.8228 recall=0.9499 f1=0.8818 support=1496

- class 1: precision=0.8882 recall=0.6608 f1=0.7578 support=902

- accuracy=0.8411

- waimai (SVM)

- class 0: precision=0.8969 recall=0.8663 f1=0.8813 support=1496

- class 1: precision=0.7901 recall=0.8348 f1=0.8119 support=902

- accuracy=0.8545

简评：

- SVM 在三套数据上均优于或匹配 NB（总体 f1/accuracy 更高），尤其在 hotel 上 SVM 大幅改善了 minority 类 recall 与 f1。

- ecommerce 数据在两模型上已接近平衡且表现良好（类间支持数大、指标稳定）。

- waimai 的 NB 对少数类 recall 较低，SVM 提升了均衡性。

## 六、代表性样例（来源：`output/labels\_\*` 的 head）

（每项：文本片段 → tokens\_join → sentiment\_score → sentiment\_label）

- hotel 示例 1:

- 文本（节选）："距离川沙公路较近,但是公交指示不对... 房间较为简单."

- tokens\_join（示例）："距离 川沙 公路 较近 , 但是 公交 指示 对 , 如果 \"\" 蔡陆线 \"\" 的话 ..."

- sentiment\_score=0.8261 → sentiment\_label=1

- ecommerce 示例 1:

- 文本（节选）："做父母一定要有刘墉这样的心态...家庭教育，真的是乐在其中."

- tokens\_join（示例）："? 做 父母 一定 要 刘墉 这样 心态 ..."

- sentiment\_score=1.0 → sentiment\_label=1

- waimai 示例 1:

- 文本："很快，好吃，味道足，量大"

- tokens\_join："很快 ， 好吃 ， 味道 足 ， 量 大"

- sentiment\_score=0.8761 → sentiment\_label=1

（注：以上示例为 head 中实际行，已确认为 UTF-8 可读中文，说明预处理写出编码问题已被修复。）

## 七、已改动的代码文件（最小化说明）

- `scripts/02\_preprocess\_and\_eda.py`（修改）

- 改动点：增强 CSV 读取回退（检测编码 -> utf-8 -> gb18030 -> latin1），使用 `on\_bad\_lines='skip'`；写出 `processed\_\*.csv` 与 `eda\_stats.csv` 时强制 `encoding='utf-8'`；保持原有的 literal-escape 解码逻辑。

- `scripts/03\_label\_and\_model.py`（已审阅，未改动）

- 该脚本已采用 utf-8 写出 labels 与报告，并包含训练集上采样与分层拆分回退策略。

## 八、重现命令（在项目根、zsh）

下面命令是在虚拟环境 `.venv` 中运行的，已被执行并成功（见运行日志）。可按需手动运行以重现：

```zsh

source .venv/bin/activate

python scripts/02\_preprocess\_and\_eda.py

python scripts/03\_label\_and\_model.py

```

或直接（合并命令）：

```zsh

.venv/bin/python3 scripts/02\_preprocess\_and\_eda.py && .venv/bin/python3 scripts/03\_label\_and\_model.py

```

## 九、我运行时的关键信息（简要日志摘要）

- 读取原始文件时，chardet 常检测到 `GB2312`，但实际以 `gb18030` 成功读取（脚本会尝试并回退）：

- "Success reading ChnSentiCorp\_htl\_all.csv with encoding=gb18030"

- 同样对 ecommerce 与 waimai 成功读取为 gb18030。

- processed CSV 与 labels/报告均已写出：例如

- `Wrote processed data to data/processed\_hotel.csv`

- `Wrote labels to output/labels\_hotel.csv`

- `Wrote classification reports for hotel` 等。

## 十、主要发现与讨论

1. 编码/转义问题已解决：此前观察到的 mojibake 与转义序列现象在最新 run 后已修复（输出文本可读）。

2. 模型表现：

- SVM 在三套数据上均为更强的基线（f1/accuracy 更高）。

- ecommerce 数据样本量大且两类支持均衡，模型稳定。

- waimai 的少数类在 NB 下 recall 较低，SVM 改进明显。

3. 评估注意事项：

- 目前各数据集的 class 支持（support）在最新拆分下都足够大（非极端 1-5 支持），因此当前 accuracy 与 f1 指标具有一定参考价值。若将来改用不同拆分/小样本测试，需注意 support 的稳定性。

## 十一、局限性

- 自动标签（SnowNLP）为弱监督/伪标签，可能带有偏差；建议将部分自动标签抽样人工复核（至少每数据集 100-300 条）以估计噪声率。

- 目前仅使用 TF-IDF + 传统分类器；后续可尝试中文预训练模型（BERT/ERNIE）或文本增强以提升泛化。

## 十二、建议的下一步（优先级）

1. 对 `labels\_\*.csv` 做抽样人工标注（200~500 条/数据集）用于估计 SnowNLP 伪标签噪声并作为少量真实标注用于微调模型。

2. 引入 Stratified K-Fold（例如 5 折），计算每折混淆矩阵与 PR 曲线，导出平均/标准差指标，增加评估稳健性（推荐优先执行）。

3. 比较不平衡处理方法（上采样、下采样、文本增强、代价敏感学习、阈值校准），并把重点放在 minority recall/precision 的改善上。

4. 若需最终提交（课程/论文），我可以把本草稿整理成 PDF 并补充图表（长度分布图/混淆矩阵/PR 曲线），也可以把关键结果写入报告 Word 格式。

## 十三、结论

已完成：预处理（含编码修复）、伪标签生成、TF-IDF + NB/SVM 训练并输出分类报告。SVM 为更稳健的基线。当前结果足够用于撰写中期实验报告；如果你希望我继续，我可以继续执行 K-Fold、生成图表并形成最终版报告（含图片与混淆矩阵）。

---

如需我把本报告转换为 PDF/Word 并把它放到 `output/` 下，或继续做 5 折 CV 并生成图表，请直接回复 "生成最终报告(PDF)" 或 "做5折CV"，我会继续执行并把结果写入工作区。

## 十四、本次新增的 CV 输出与抽样（已生成）

我已对三个数据集执行 5 折分层交叉验证，并生成每折的 PR/ROC 曲线与 SVM 混淆矩阵，以及每数据集的平均/标准差指标文件。关键文件：

- `output/cv/{hotel,ecommerce,waimai}/metrics\_summary.csv`（每个数据集，包含 precision/recall/f1/accuracy 的 mean/std）

- `output/cv/{name}/metrics\_per\_fold.csv`（每折的原始度量）

- 各折图像示例（保存在对应文件夹）：

- `fold\_{i}\_svm\_confusion.png`

- `fold\_{i}\_svm\_pr.png`

- `fold\_{i}\_svm\_roc.png`

此外，为每个数据集生成了供人工复核的抽样文件（最多 200 条，分层抽样）：

- `output/samples\_for\_annotation\_hotel.csv`

- `output/samples\_for\_annotation\_ecommerce.csv`

- `output/samples\_for\_annotation\_waimai.csv`

下面为 CV 汇总（从 `metrics\_summary.csv` 摘要）：

- hotel (nb): precision\_mean=0.961 ±0.012, recall\_mean=0.427 ±0.0099, f1\_mean=0.591 ±0.0081, accuracy\_mean=0.774 ±0.0026

- hotel (svm): precision\_mean=0.807 ±0.026, recall\_mean=0.844 ±0.014, f1\_mean=0.825 ±0.018, accuracy\_mean=0.862 ±0.016

- ecommerce (nb): precision\_mean=0.886 ±0.0032, recall\_mean=0.855 ±0.0043, f1\_mean=0.870 ±0.0037, accuracy\_mean=0.878 ±0.0034

- ecommerce (svm): precision\_mean=0.899 ±0.0019, recall\_mean=0.901 ±0.0058, f1\_mean=0.900 ±0.0028, accuracy\_mean=0.904 ±0.0024

- waimai (nb): precision\_mean=0.855 ±0.0136, recall\_mean=0.629 ±0.0223, f1\_mean=0.725 ±0.0173, accuracy\_mean=0.820 ±0.0096

- waimai (svm): precision\_mean=0.788 ±0.0178, recall\_mean=0.802 ±0.0169, f1\_mean=0.795 ±0.0108, accuracy\_mean=0.844 ±0.0088

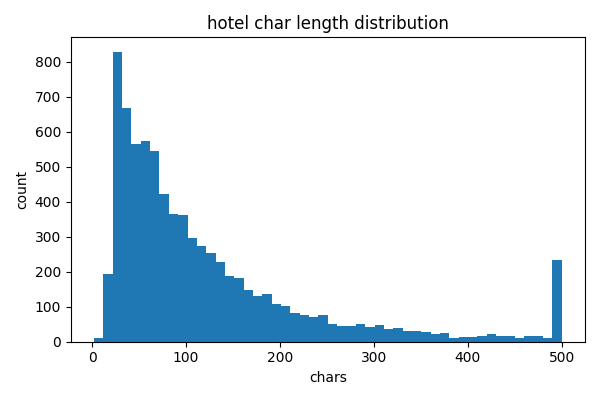
这些结果将被包含在最终 PDF 中，并附上每折的混淆矩阵与 PR/ROC 图像作为附录。

## 十五：现在我要把此 Markdown 转成 PDF 并保存为 `output/experiment\_report.pdf`（几种可能的实现路径）

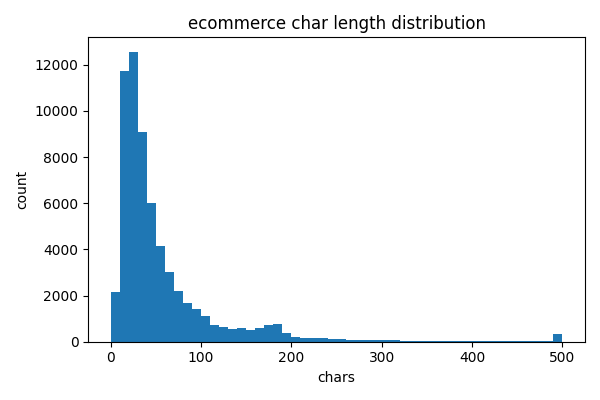
我将先尝试使用系统上的 `pandoc` 将 Markdown 转为 PDF（优选 xelatex 字体支持中文）；若系统无 pandoc 或无 LaTeX，我会回退到在虚拟环境中用 Python 库生成一个包含文本和关键图片的基础 PDF（保证可阅读）。

现在我将检测系统上是否有 `pandoc` 或 `wkhtmltopdf`，以决定转换方法。后续我会运行转换并把最终 `output/experiment\_report.pdf` 放到工作区。

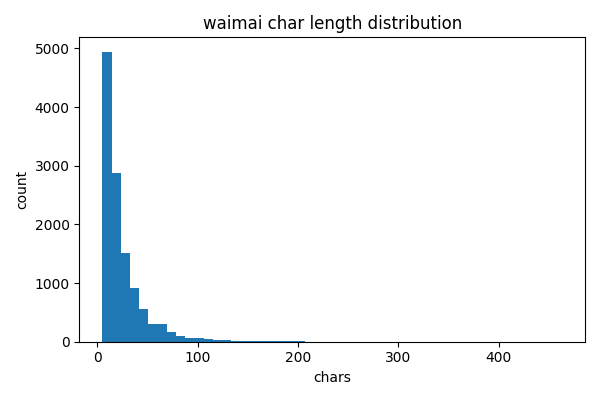
## Length histogram: length\_hist\_hotel.png



## Length histogram: length\_hist\_ecommerce.png



## Length histogram: length\_hist\_waimai.png



# Cross-validation results: hotel

Summary (mean ± std) for precision/recall/f1/accuracy:

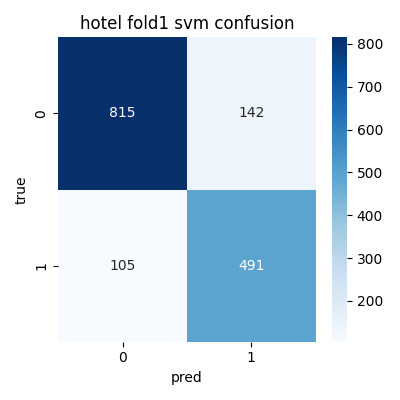
|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| dataset | model | precision\_mean | precision\_std | recall\_mean | recall\_std | f1\_mean | f1\_std | accuracy\_mean | accuracy\_std |
| hotel | nb | 0.9609653422765791 | 0.0120655246160896 | 0.4270779989848289 | 0.0098815485405413 | 0.5912391363500674 | 0.0080840915880101 | 0.7737282678686414 | 0.0026392661341814 |
| hotel | svm | 0.8066876794014657 | 0.0261944027922784 | 0.8440928317635779 | 0.0140747189531787 | 0.8248328911064584 | 0.0179796390733358 | 0.8624597553122987 | 0.0156114533320901 |

Summary:

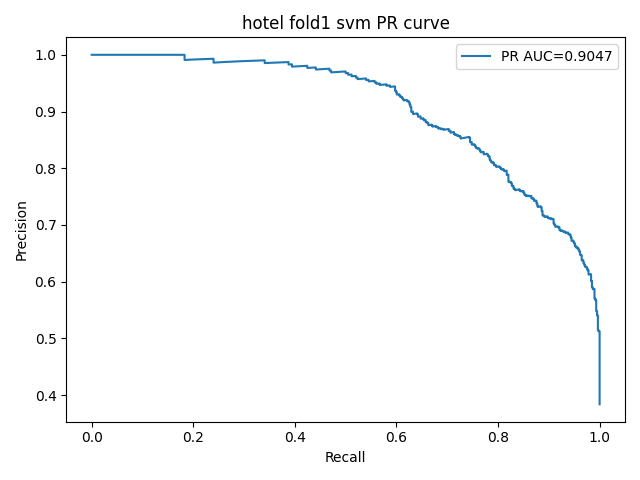
nb: f1\_mean=0.591 ± 0.008  
svm: f1\_mean=0.825 ± 0.018

## Fold 1

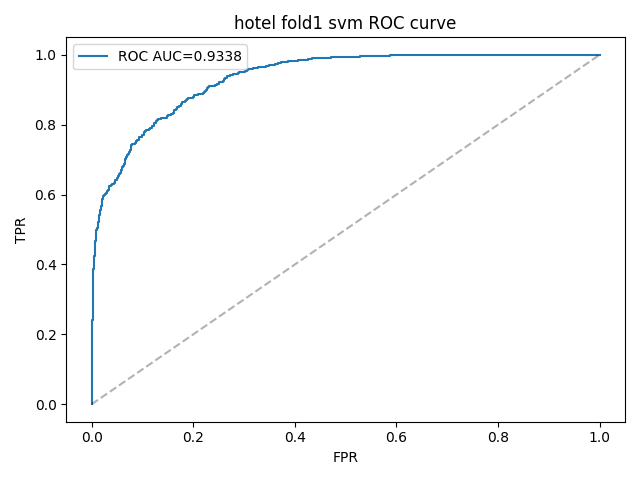
SVM confusion matrix (rows=true labels, cols=predicted labels):



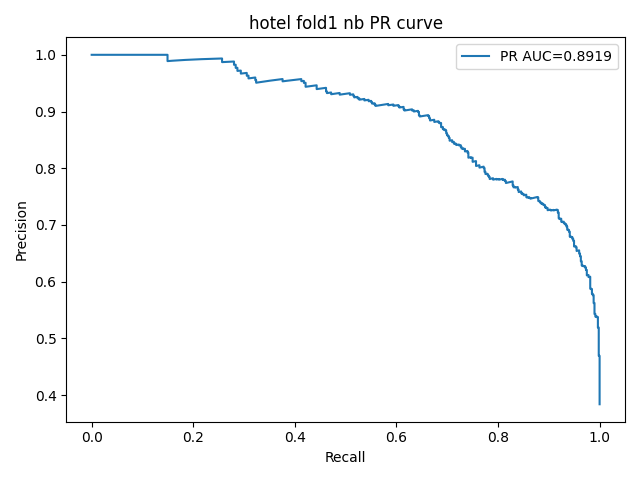
SVM PR curve (fold 1):



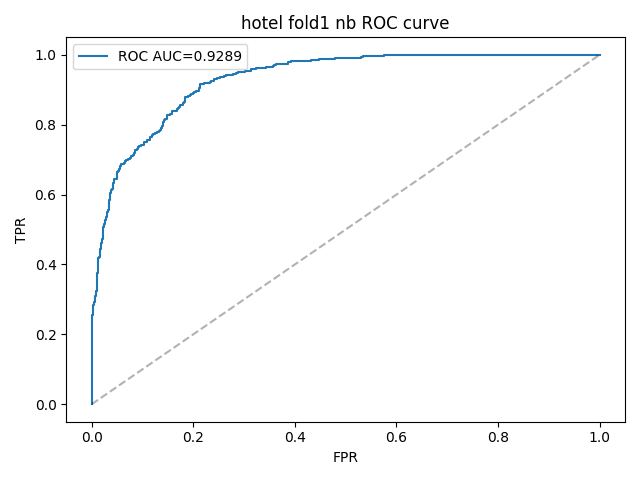
SVM ROC curve (fold 1):



NB PR curve (fold 1):

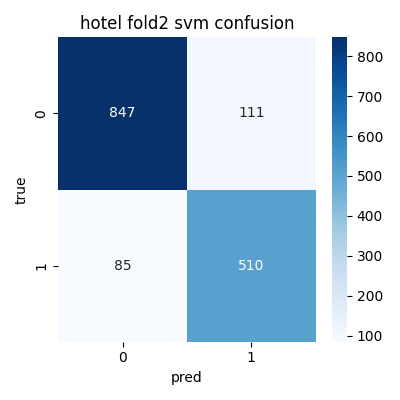


NB ROC curve (fold 1):

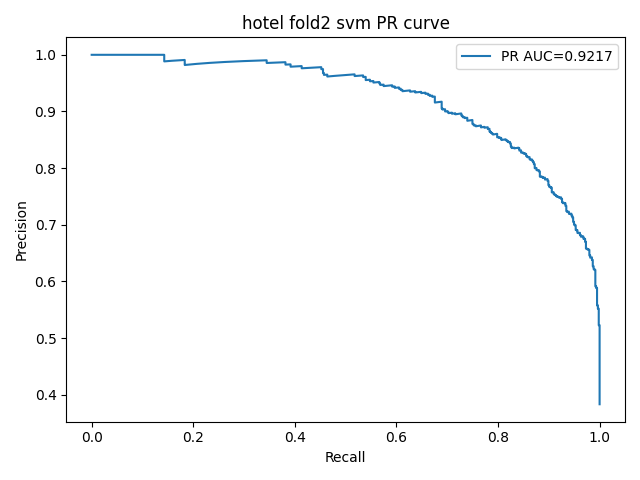


## Fold 2

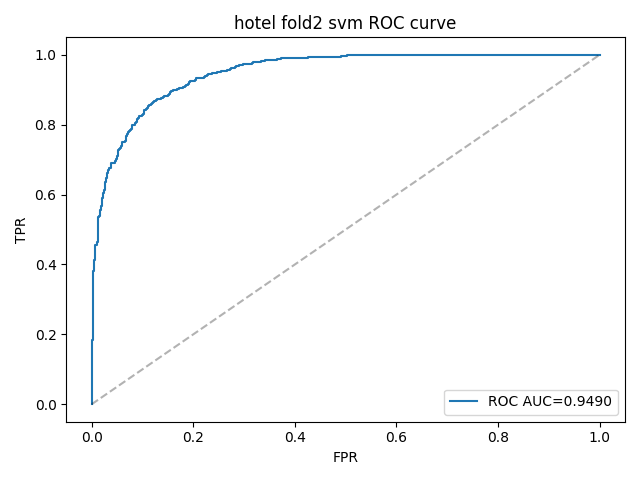
SVM confusion matrix (rows=true labels, cols=predicted labels):



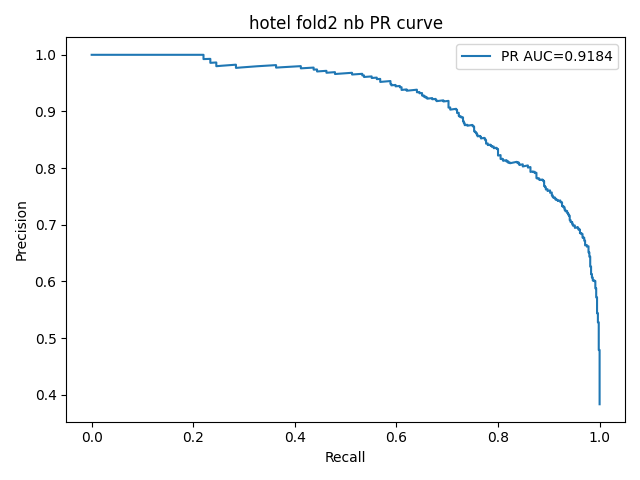
SVM PR curve (fold 2):



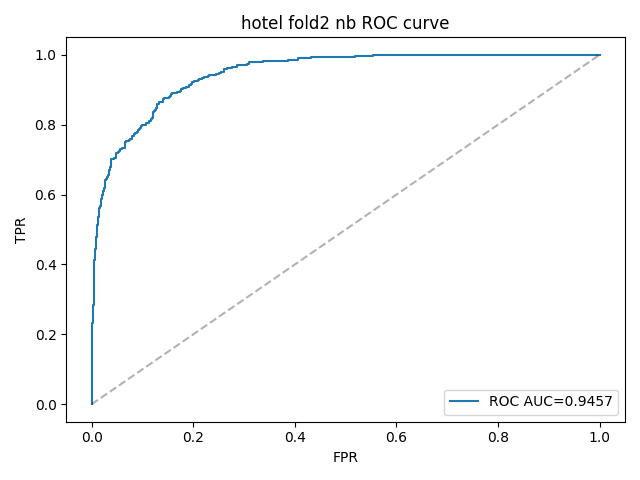
SVM ROC curve (fold 2):



NB PR curve (fold 2):

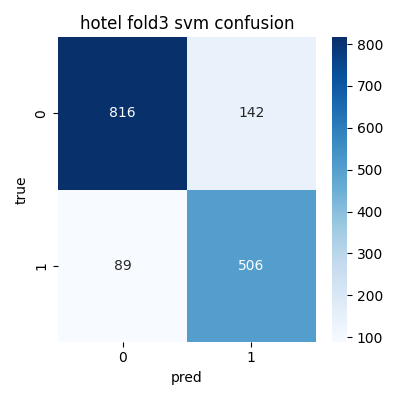


NB ROC curve (fold 2):

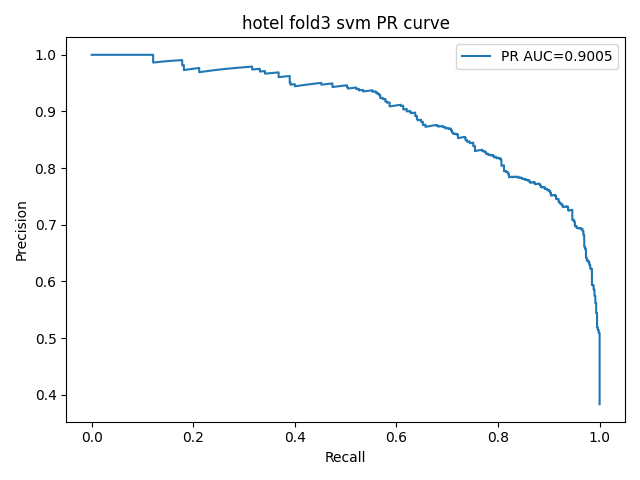


## Fold 3

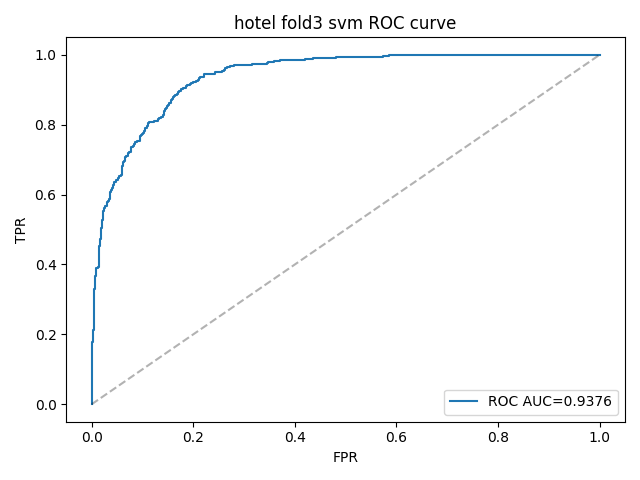
SVM confusion matrix (rows=true labels, cols=predicted labels):



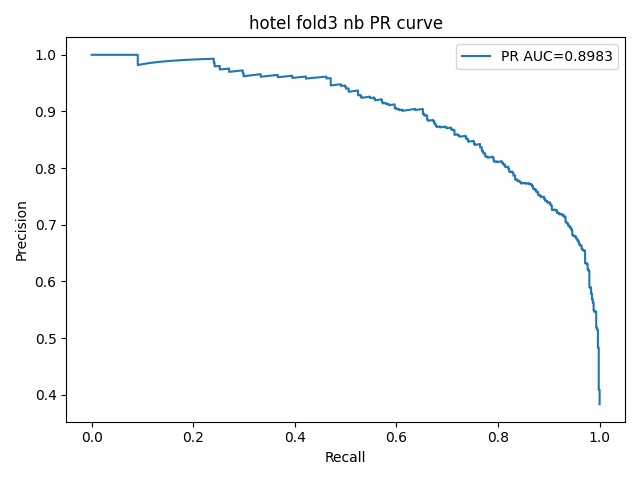
SVM PR curve (fold 3):



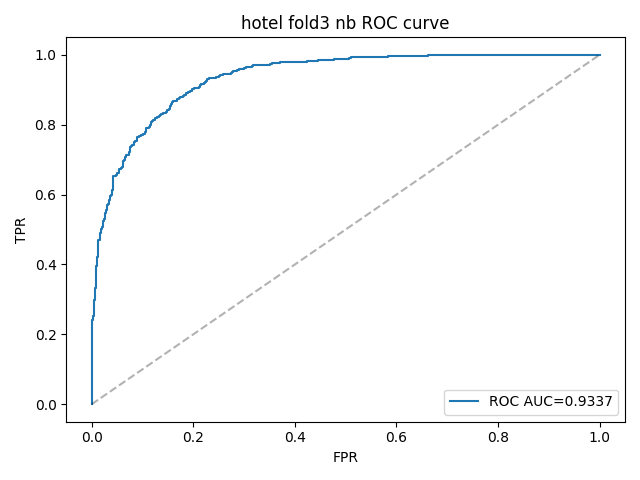
SVM ROC curve (fold 3):



NB PR curve (fold 3):

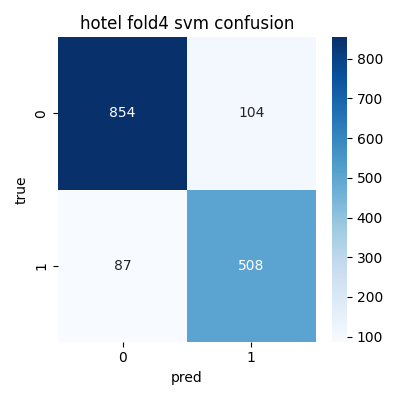


NB ROC curve (fold 3):

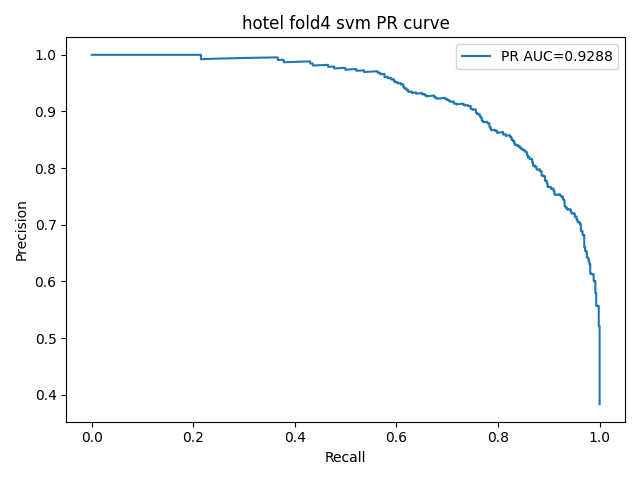


## Fold 4

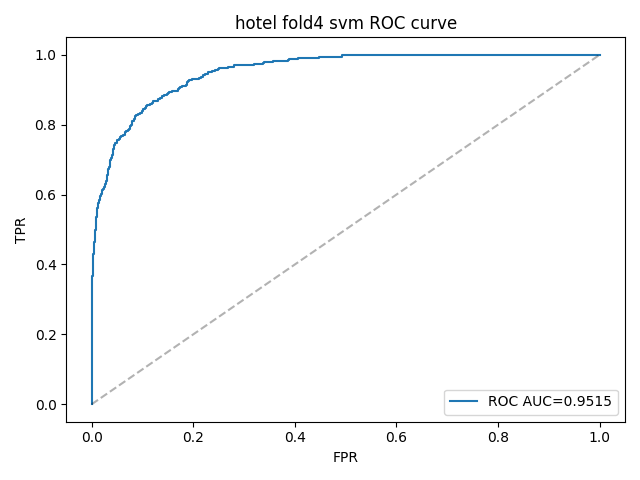
SVM confusion matrix (rows=true labels, cols=predicted labels):



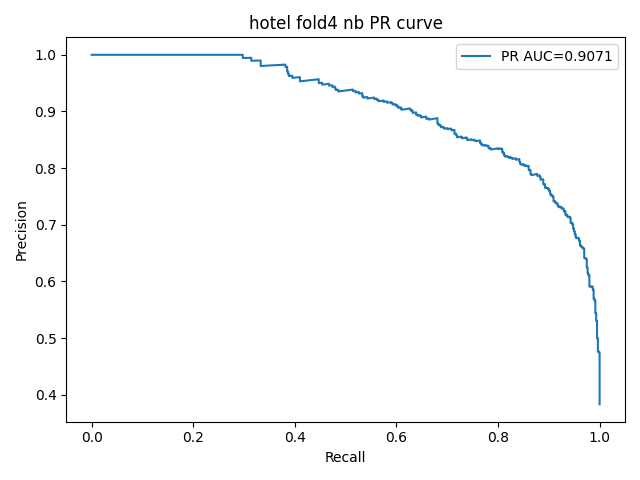
SVM PR curve (fold 4):



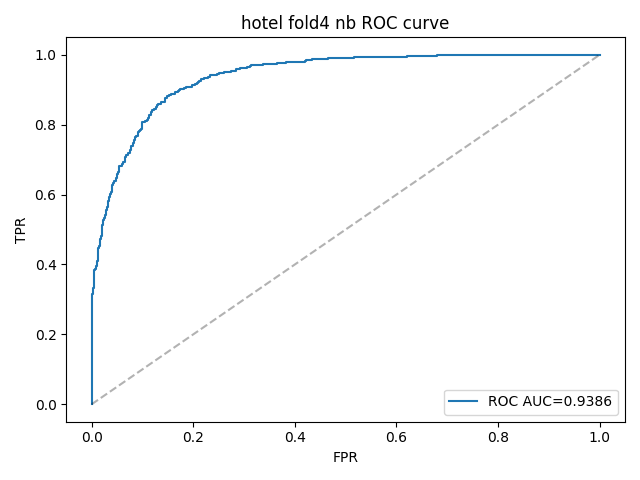
SVM ROC curve (fold 4):



NB PR curve (fold 4):

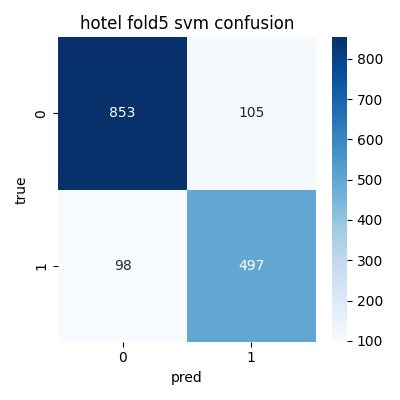


NB ROC curve (fold 4):

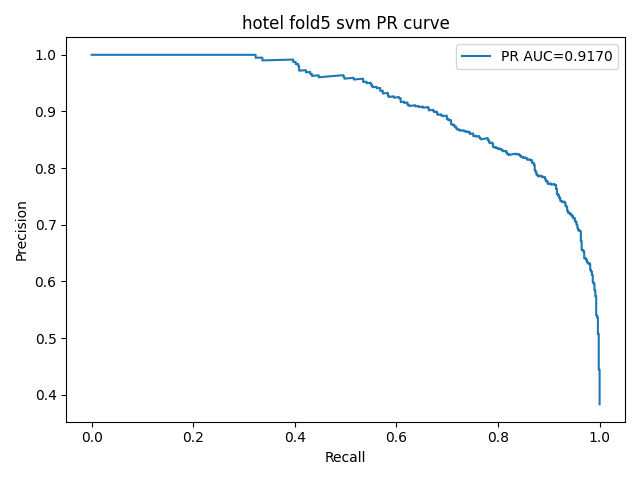


## Fold 5

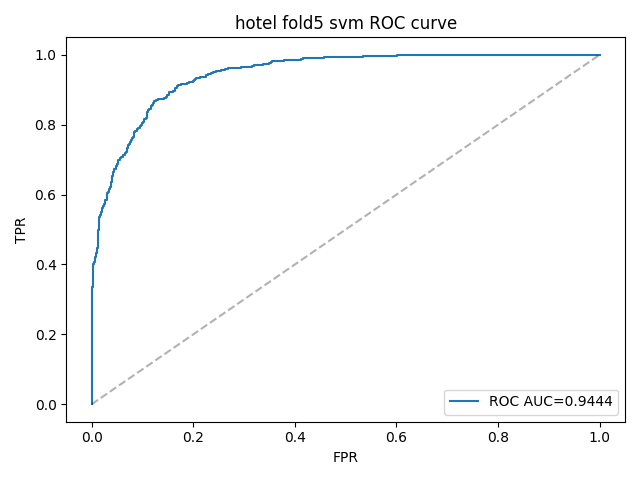
SVM confusion matrix (rows=true labels, cols=predicted labels):



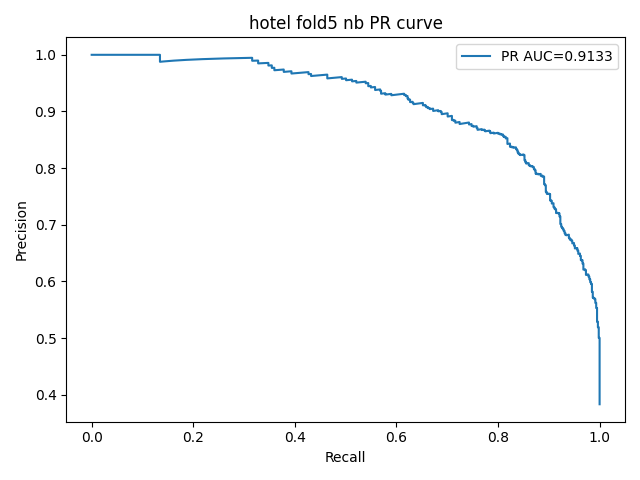
SVM PR curve (fold 5):



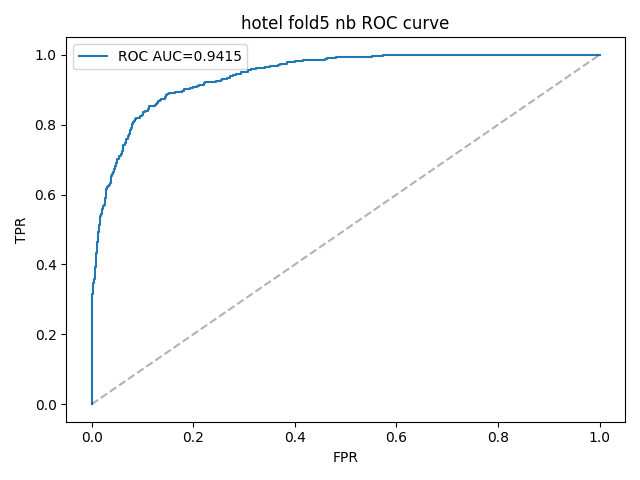
SVM ROC curve (fold 5):



NB PR curve (fold 5):



NB ROC curve (fold 5):



# Cross-validation results: ecommerce

Summary (mean ± std) for precision/recall/f1/accuracy:

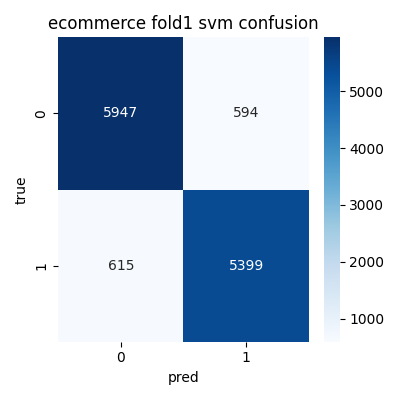
|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| dataset | model | precision\_mean | precision\_std | recall\_mean | recall\_std | f1\_mean | f1\_std | accuracy\_mean | accuracy\_std |
| ecommerce | nb | 0.8860960764953221 | 0.0032102462032703 | 0.8553042899900232 | 0.0043158848593788 | 0.8704268414794797 | 0.0036732757036796 | 0.8780227626133399 | 0.0033713617905998 |
| ecommerce | svm | 0.8992586462016808 | 0.0018500589479147 | 0.9011972065181244 | 0.0057856858329913 | 0.9002175761437566 | 0.0028170947692969 | 0.9043075581349976 | 0.0024260418887891 |

Summary:

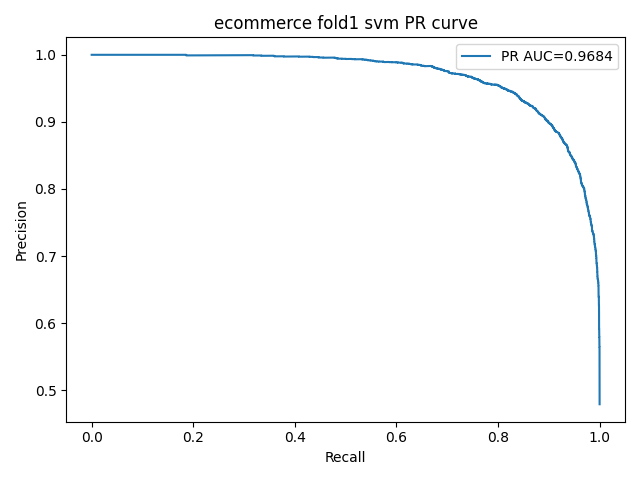
nb: f1\_mean=0.870 ± 0.004  
svm: f1\_mean=0.900 ± 0.003

## Fold 1

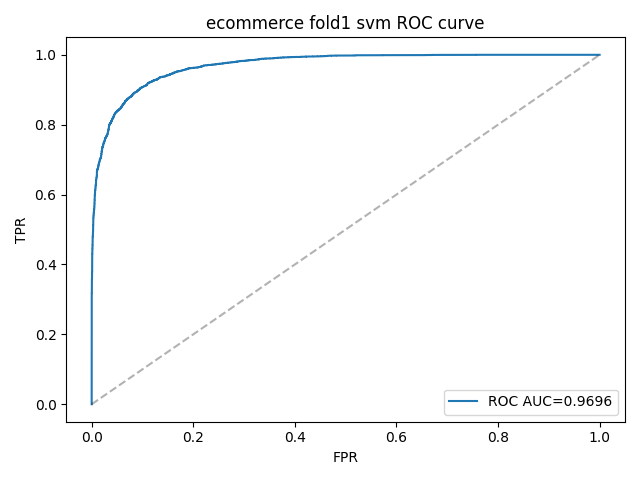
SVM confusion matrix (rows=true labels, cols=predicted labels):



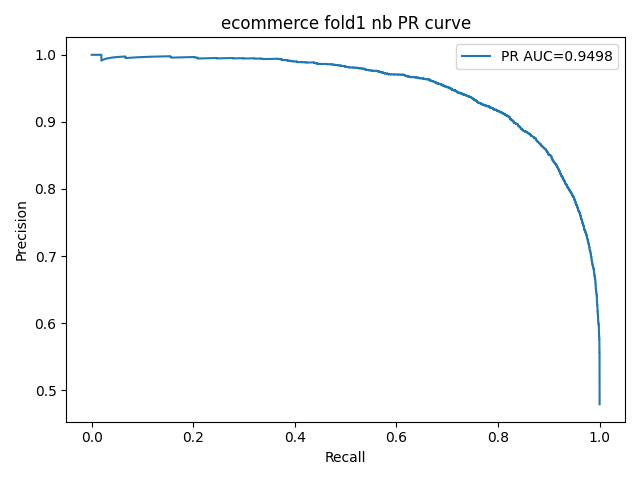
SVM PR curve (fold 1):



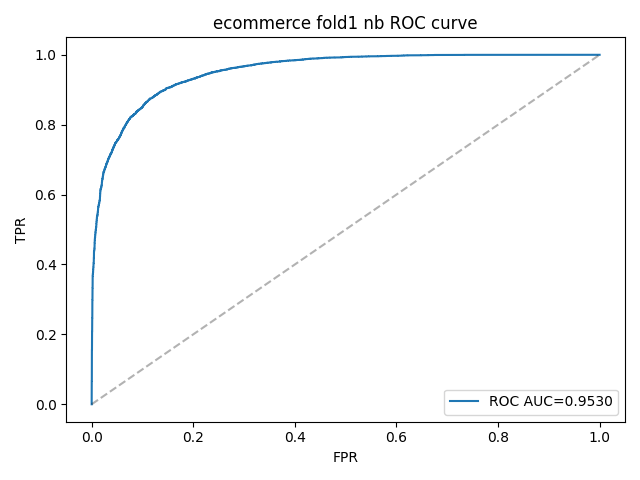
SVM ROC curve (fold 1):



NB PR curve (fold 1):

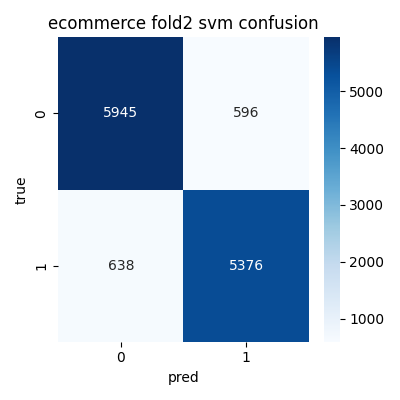


NB ROC curve (fold 1):

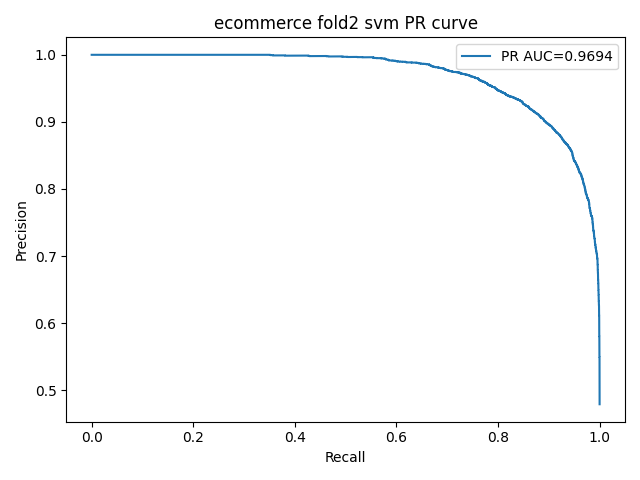


## Fold 2

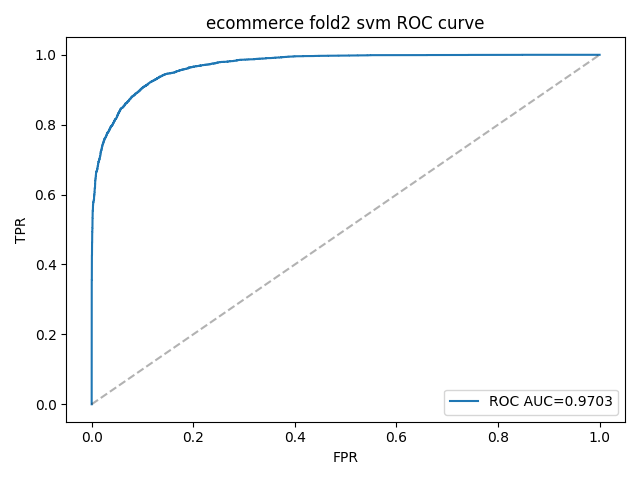
SVM confusion matrix (rows=true labels, cols=predicted labels):



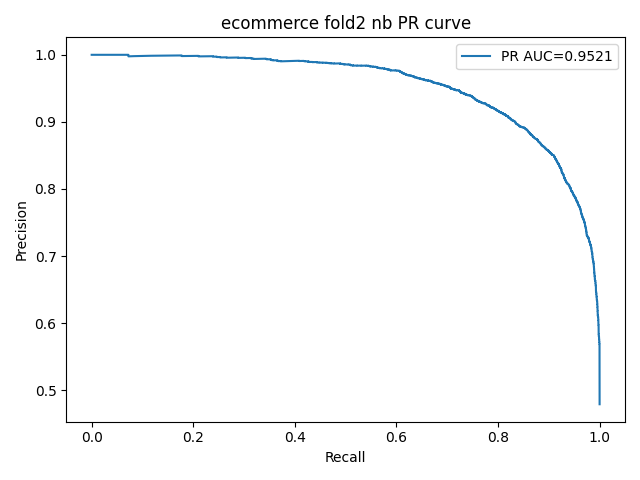
SVM PR curve (fold 2):



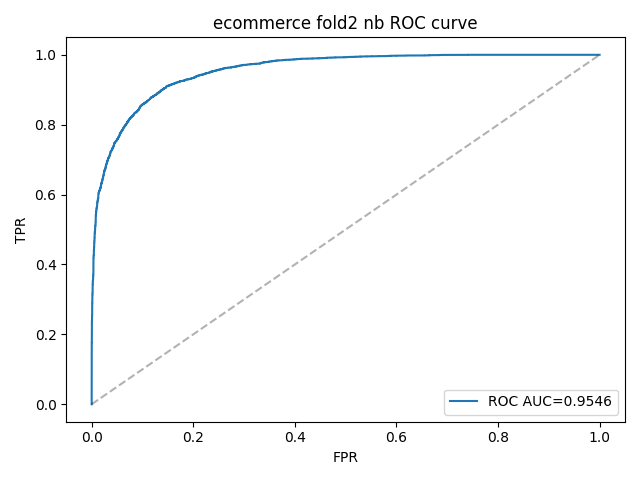
SVM ROC curve (fold 2):



NB PR curve (fold 2):

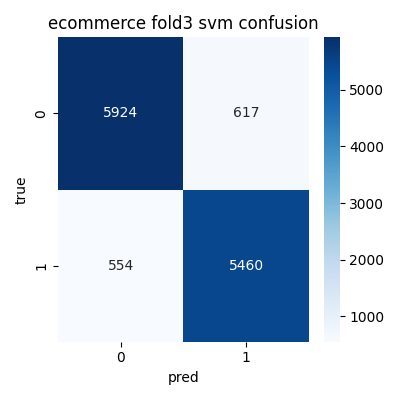


NB ROC curve (fold 2):

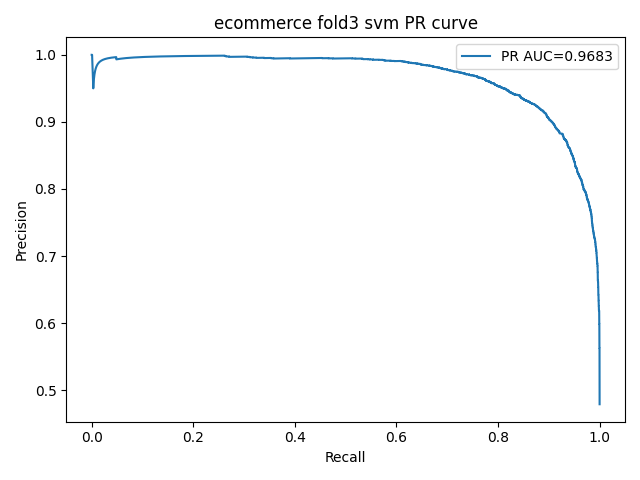


## Fold 3

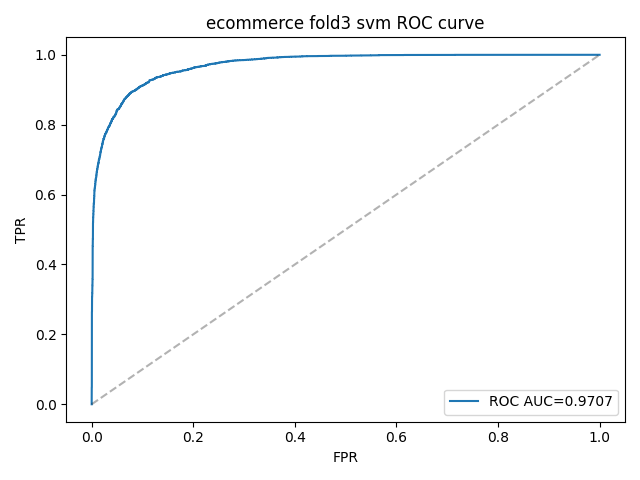
SVM confusion matrix (rows=true labels, cols=predicted labels):



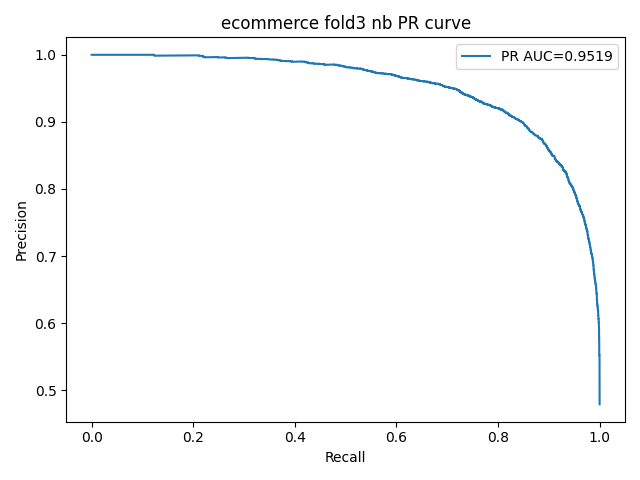
SVM PR curve (fold 3):



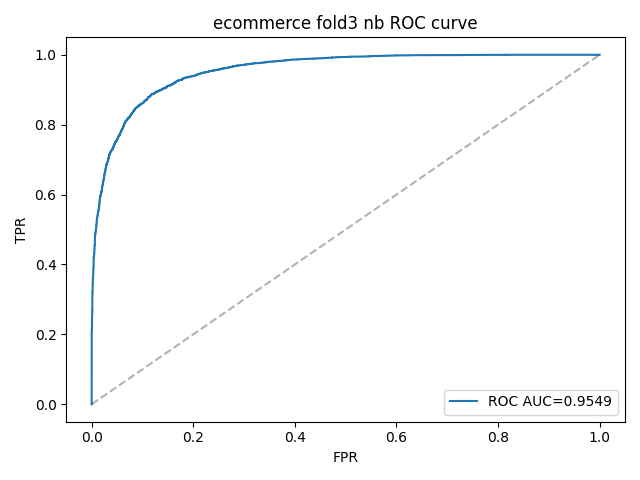
SVM ROC curve (fold 3):



NB PR curve (fold 3):

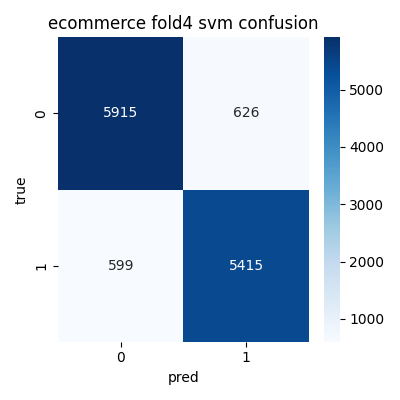


NB ROC curve (fold 3):

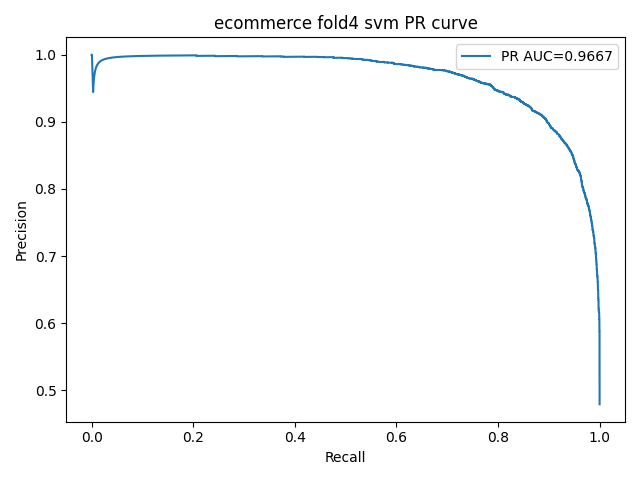


## Fold 4

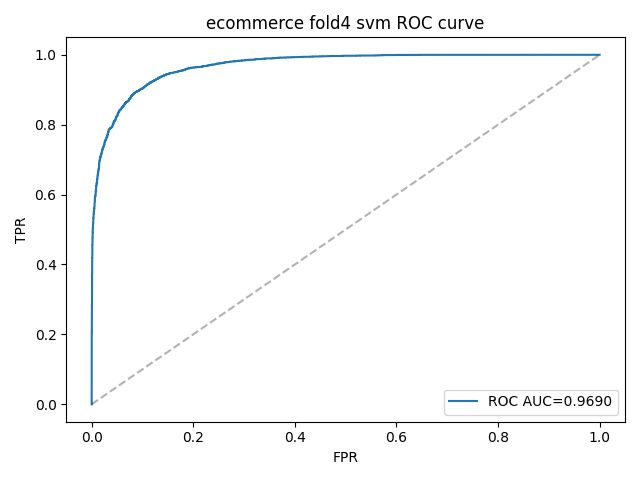
SVM confusion matrix (rows=true labels, cols=predicted labels):



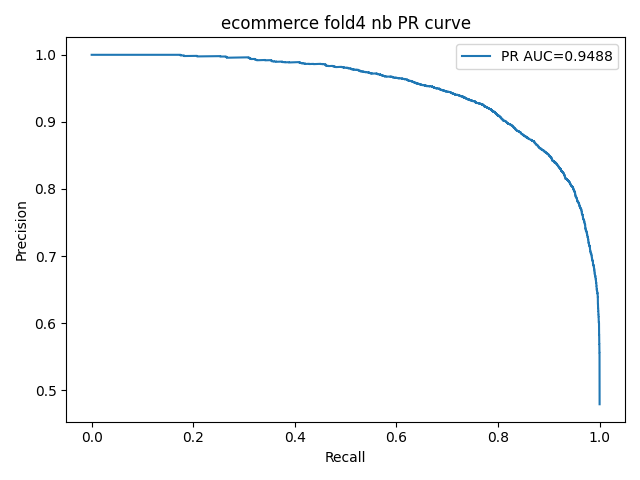
SVM PR curve (fold 4):



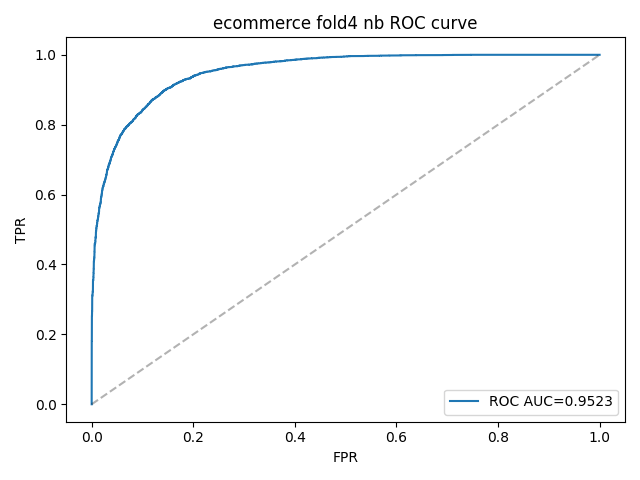
SVM ROC curve (fold 4):



NB PR curve (fold 4):

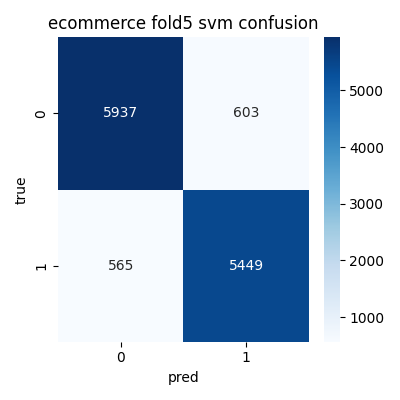


NB ROC curve (fold 4):

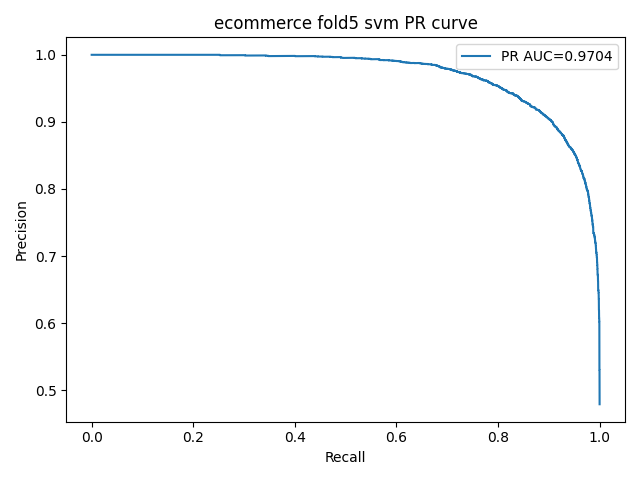


## Fold 5

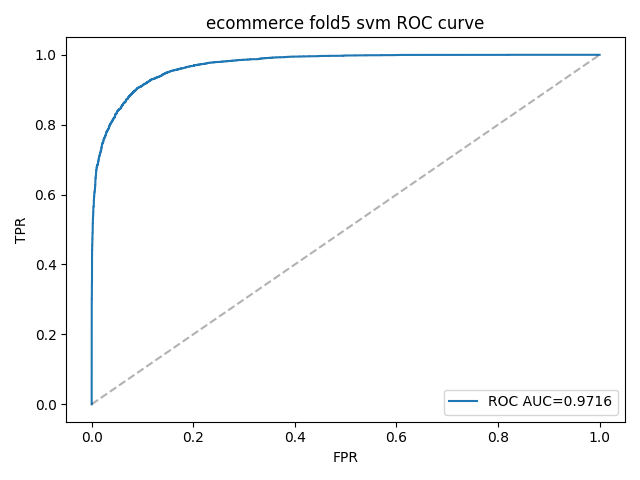
SVM confusion matrix (rows=true labels, cols=predicted labels):



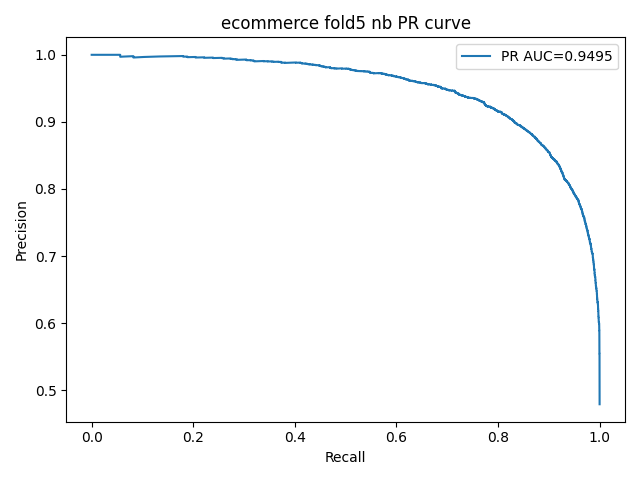
SVM PR curve (fold 5):



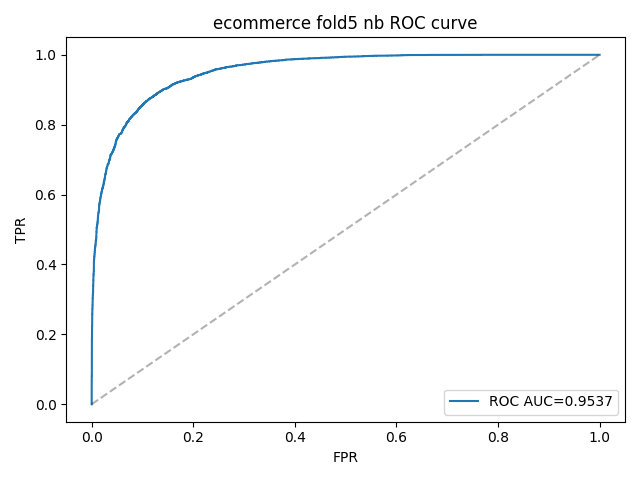
SVM ROC curve (fold 5):



NB PR curve (fold 5):



NB ROC curve (fold 5):



# Cross-validation results: waimai

Summary (mean ± std) for precision/recall/f1/accuracy:

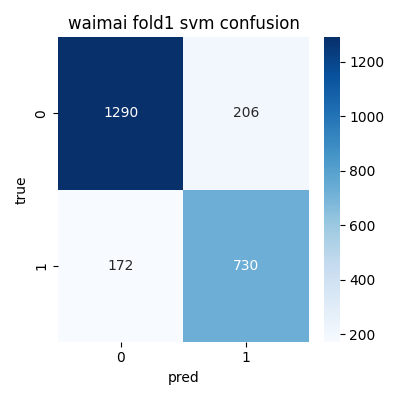
|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| dataset | model | precision\_mean | precision\_std | recall\_mean | recall\_std | f1\_mean | f1\_std | accuracy\_mean | accuracy\_std |
| waimai | nb | 0.85478053393899 | 0.0136486262788963 | 0.6291804376019746 | 0.0222506608675548 | 0.7246873569889869 | 0.0173014991270013 | 0.8203039106083049 | 0.0095741837828314 |
| waimai | svm | 0.7880158399618473 | 0.017812085702148 | 0.8019480695260993 | 0.0169173739058279 | 0.7947368877577423 | 0.0108365350994533 | 0.8441636630163574 | 0.0088353062976555 |

Summary:

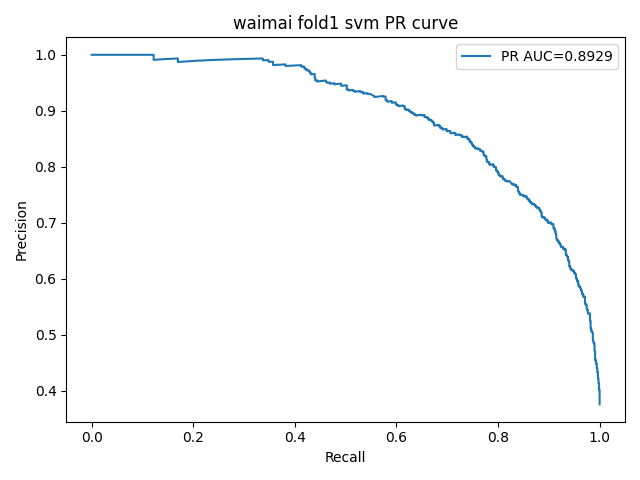
nb: f1\_mean=0.725 ± 0.017  
svm: f1\_mean=0.795 ± 0.011

## Fold 1

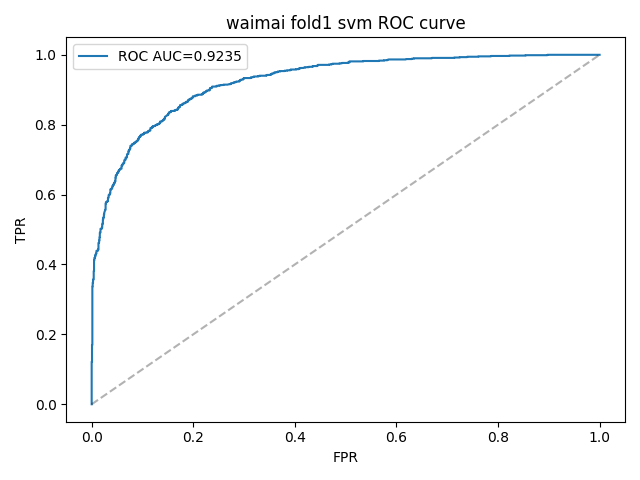
SVM confusion matrix (rows=true labels, cols=predicted labels):



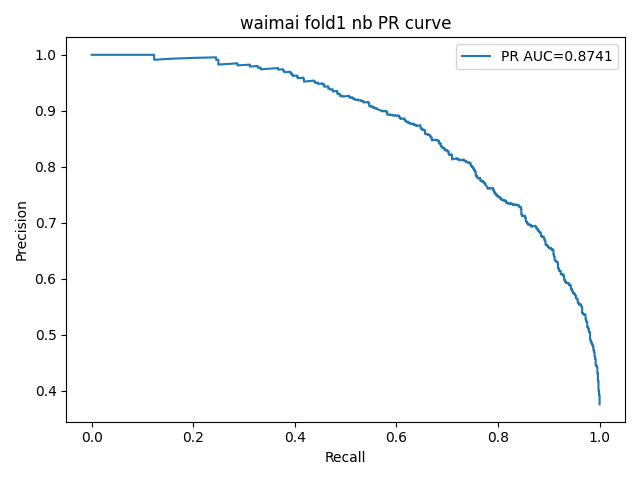
SVM PR curve (fold 1):



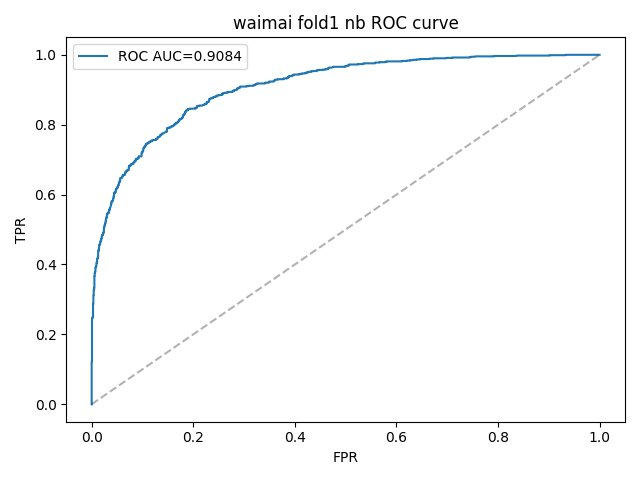
SVM ROC curve (fold 1):



NB PR curve (fold 1):

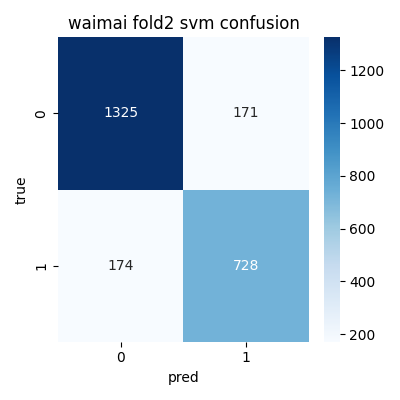


NB ROC curve (fold 1):

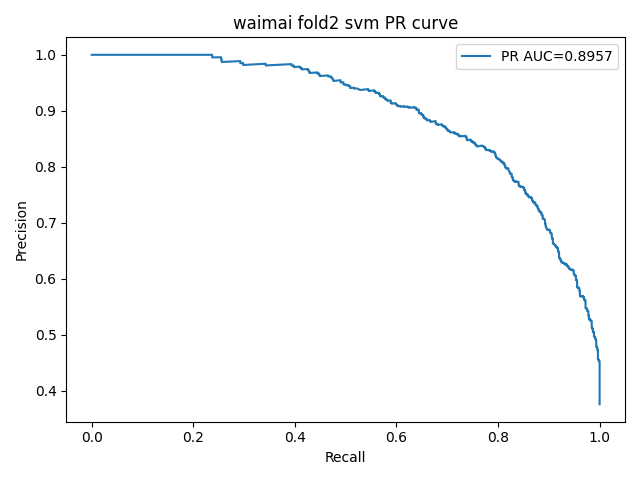


## Fold 2

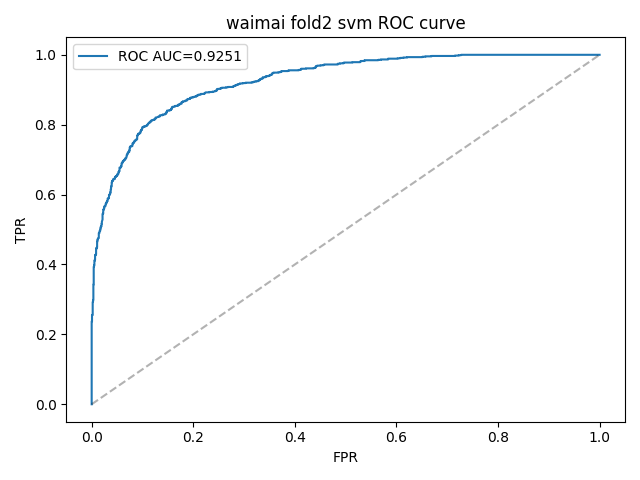
SVM confusion matrix (rows=true labels, cols=predicted labels):



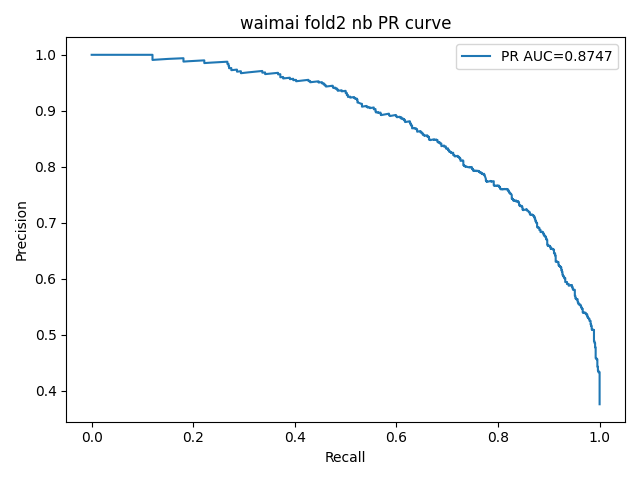
SVM PR curve (fold 2):



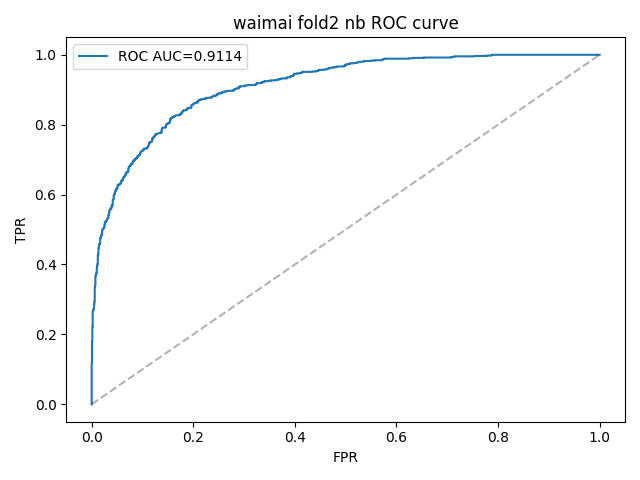
SVM ROC curve (fold 2):



NB PR curve (fold 2):

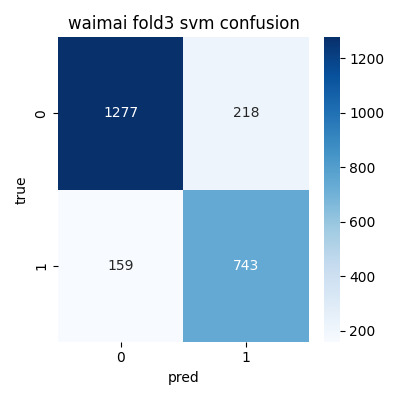


NB ROC curve (fold 2):

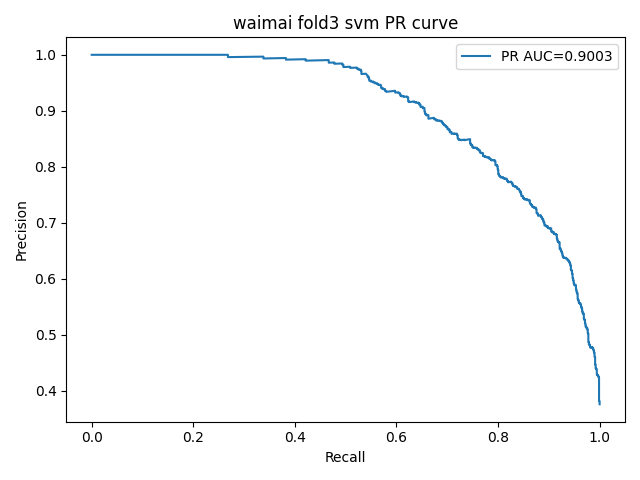


## Fold 3

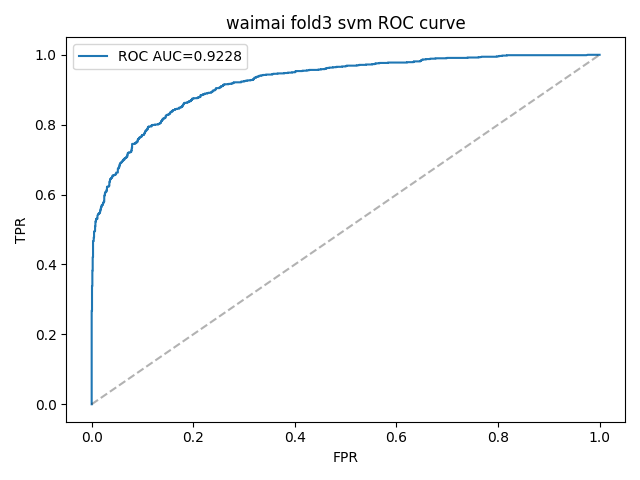
SVM confusion matrix (rows=true labels, cols=predicted labels):



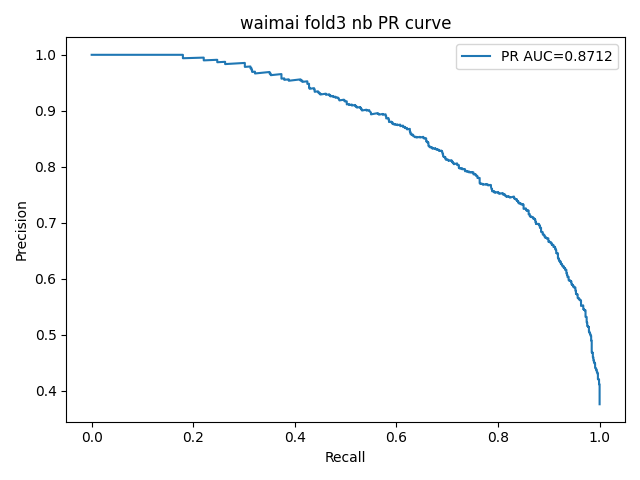
SVM PR curve (fold 3):



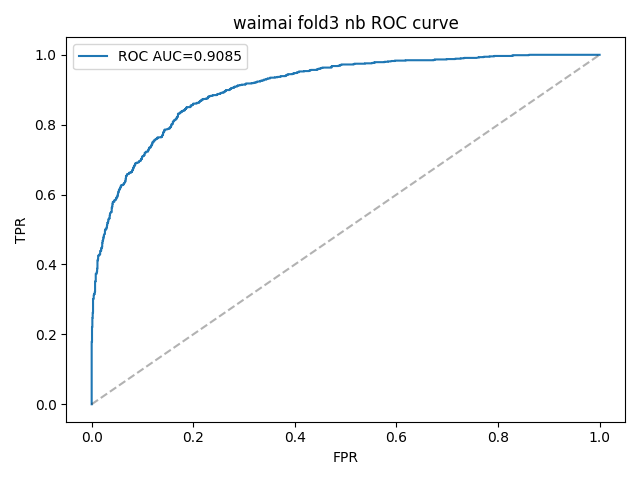
SVM ROC curve (fold 3):



NB PR curve (fold 3):

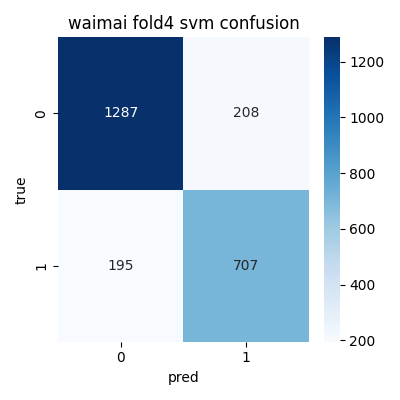


NB ROC curve (fold 3):

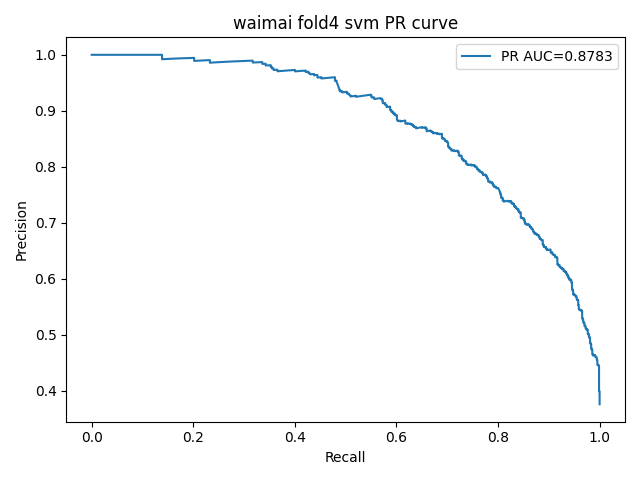


## Fold 4

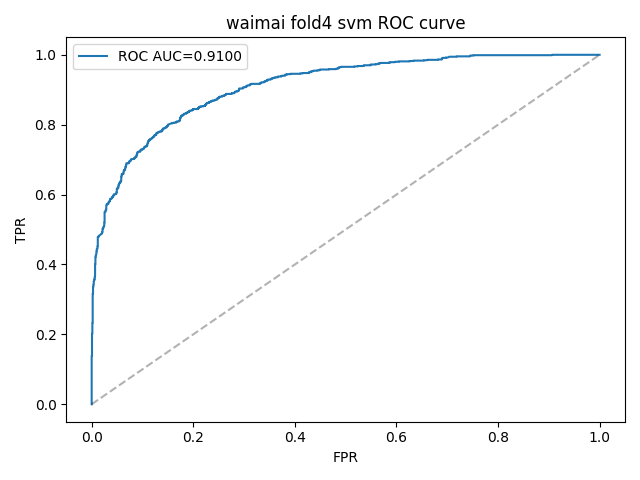
SVM confusion matrix (rows=true labels, cols=predicted labels):



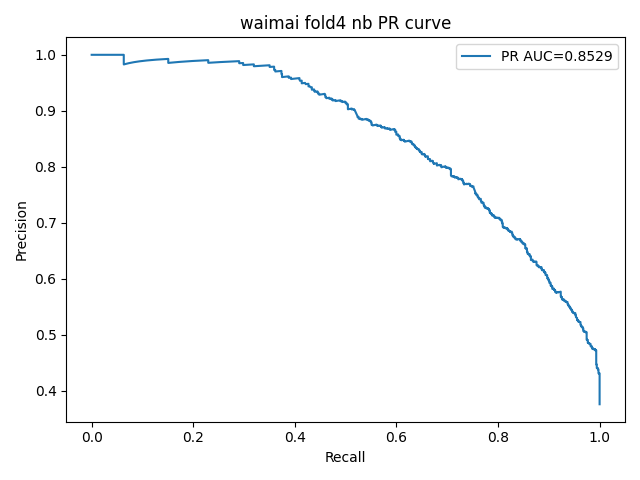
SVM PR curve (fold 4):



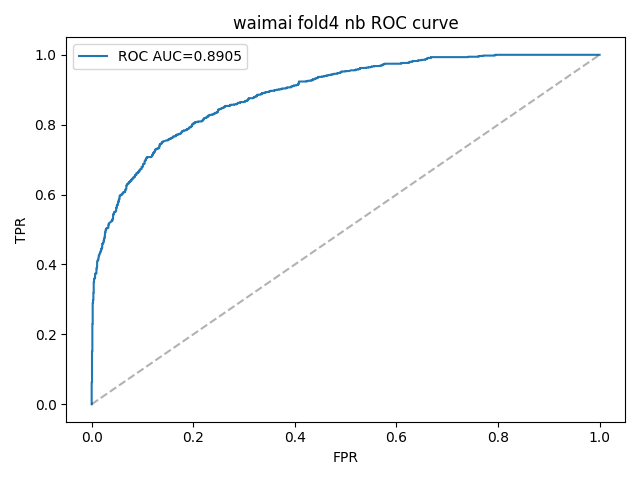
SVM ROC curve (fold 4):



NB PR curve (fold 4):

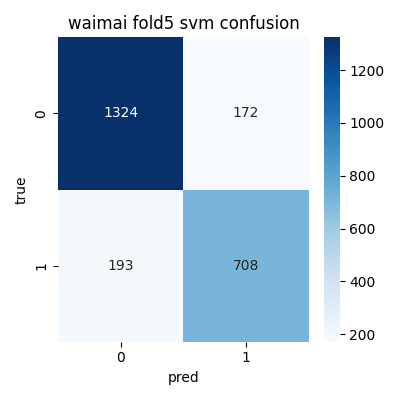


NB ROC curve (fold 4):

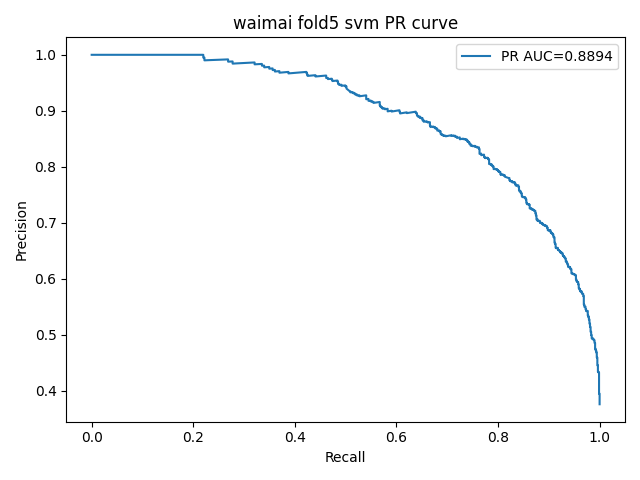


## Fold 5

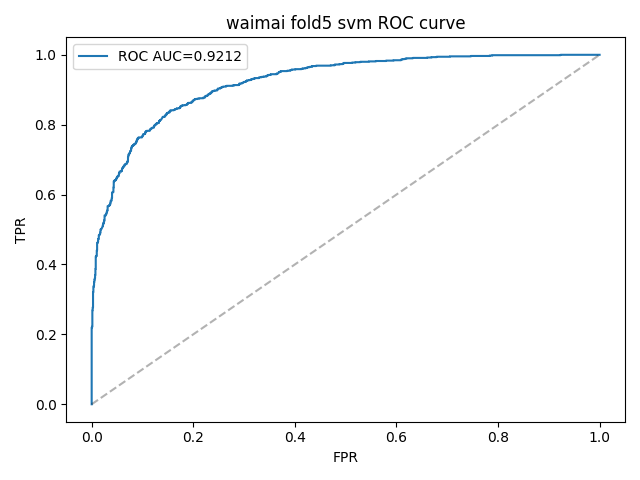
SVM confusion matrix (rows=true labels, cols=predicted labels):



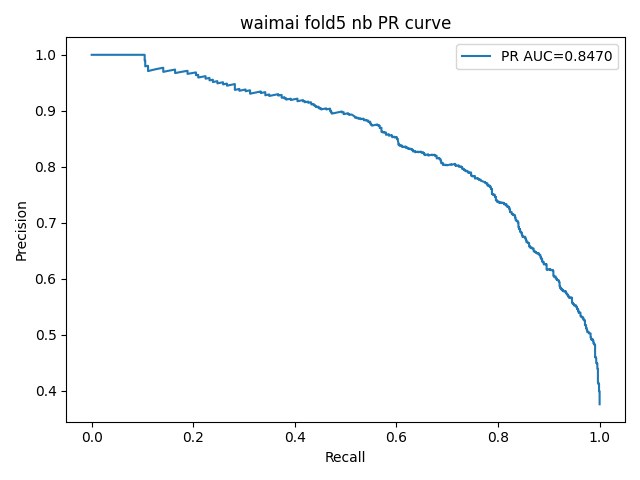
SVM PR curve (fold 5):



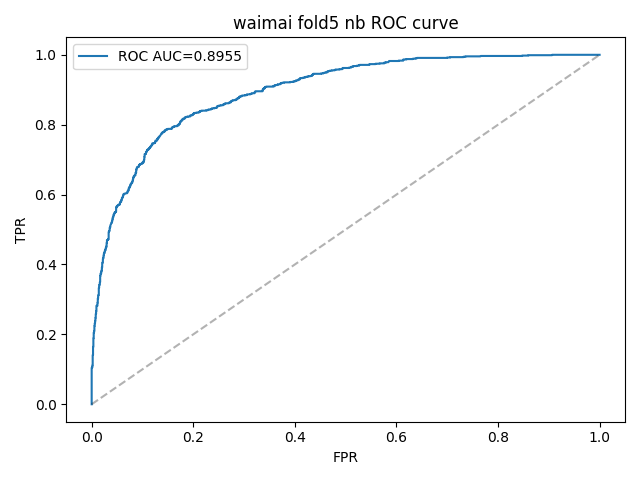
SVM ROC curve (fold 5):



NB PR curve (fold 5):



NB ROC curve (fold 5):



## Sample for annotation: samples\_for\_annotation\_hotel.csv

review,tokens\_join,sentiment\_score,sentiment\_label

方便倒是真方便。可是晚上噪音太大了。正好是个十字路口。整个晚上全过车。鸣笛不止。睡觉轻的还是不要去住了,方便 倒 真 方便 。 可是 晚上 噪音 太大 。 正好 个 十字路口 。 整个 晚上 全过 车 。 鸣笛 不止 。 睡觉 轻 还是 不要 去 住,0.0682378082959145,0

很精致的藏式酒店，坐落在松赞林寺旁边，早朝松赞林寺再方便不过；酒店服务挺周到，房间很干净，被褥十分舒适；餐厅不大但很精巧，最喜欢的就是没有饭菜油烟的味道，因为饭菜都是在楼下厨房做好从一个窗口送进来的。餐厅的价格很合理，口味也不错；早餐品种虽然不多，不过藏式粑粑和牦牛酸奶不可不尝；招牌的藏式火锅完全没有想象中的浓油重盐，非常清淡可口，尤其是里面的豆腐猪肉丸，很好吃！不喜欢清淡对口味的可以跟卓玛要个蘸水；喝不完的酥油茶还可以留到第二天，卓玛会热了端过来；酒店大堂旁边的精品柜台出售的藏饰做工非常精美，是经理专程从尼泊尔挑回来的，完全不是外面那些粗制滥造的恶俗款式，当然价格也是不菲，找经理也只能有9折的优惠。酒店不好的地方就是门口那条路还没修好，离县城稍微有点距离，没有包车的要当心出租车司机漫天要价。酒店不提供一次性拖鞋包括洗漱用品，要自己带齐哦！补充点评2007年11月1日：网友说的看星星的房间我觉得如果是这个季节入住的话其实不需要的，有阳台的话会比较冷，大冷天我想没人愿意在外面吹北风吧；夏天的话还是可以的，到前台的时候可以提出要求，服务员会耐心的打开房间让你看的,精致 藏式 酒店 ， 坐落 松 赞林寺 旁边 ， 早朝松 赞林寺 再 方便 不过 ； 酒店 服务 挺 周到 ， 房间 干净 ， 被褥 十分 舒适 ； 餐厅 不大 但 精巧 ， 最 喜欢 就是 没有 饭菜 油烟 味道 ， 因为 饭菜 楼下 厨房 做好 从 一个 窗口 送进来 。 餐厅 价格 合理 ， 口味 不错 ； 早餐 品种 虽然 多 ， 不过 藏式 粑 粑 牦牛 酸奶 不可不 尝 ； 招牌 藏式 火锅 完全 没有 想象 中 浓油 重盐 ， 非常 清淡 可口 ， 尤其 里面 豆腐 猪肉 丸 ， 好吃 ！ 喜欢 清淡 对 口味 可以 跟 卓玛要 个 蘸 水 ； 喝 完 酥油茶 还 可以 留 到 第二天 ， 卓玛会 热 端过来 ； 酒店 大堂 旁边 精品 柜台 出售 藏饰 做工 非常 精美 ， 经理 专程 从 尼泊尔 挑 回来 ， 完全 不是 外面 那些 粗制滥造 恶俗 款式 ， 当然 价格 不菲 ， 找 经理 只能 9 折 优惠 。 酒店 不好 地方 就是 门口 那条 路 还 没 修好 ， 离 县城 稍微 有点 距离 ， 没有 包车 要 当心 出租车 司机 漫天要价 。 酒店 提供 一次性 拖鞋 包括 洗漱 用品 ， 要 自己 带齐 哦 ！ 补充 点评 2007 年 11 月 1 日 ： 网友 说 看 星星 房间 觉得 如果 这个 季节 入住 的话 其实 需要 ， 阳台 的话 会 比较 冷 ， 大 冷天 我想 没人 愿意 外面 吹 北风 吧 ； 夏天 的话 还是 可以 ， 到 前台 时候 可以 提出 要求 ， 服务员 会 耐心 打开 房间 让 你 看,6.465975843639171e-08,0

"整体尚可.优点很多就不说了.单就缺点罗列如下,望改进:1,房内夜间有异响.象是水滴或是气流的声音.2,宽带要服务台开通才可使用.开始不明白,耽误了好多事.3,被子感觉很不好,不透气,象一层塑料罩着一样,很难受(个人感受).4,落地灯插座没电.只好趴在桌下找桌后的插座.5,卫生间隔音不是一般的差.","整体 尚可 . 优点 很多 说 . 单 缺点 罗列 如下 , 望 改进 : 1 , 房 内 夜间 异响 . 象是 水滴 或是 气流 声音 . 2 , 宽带 要 服务台 开通 才 可 使用 . 开始 明白 , 耽误 好多 事 . 3 , 被子 感觉 不好 , 不透气 , 象 一层 塑料 罩 着 一样 , 难受 ( 个人 感受 ) . 4 , 落地灯 插座 没电 . 只好 趴在 桌下 找 桌后 插座 . 5 , 卫生间 隔音 不是 一般 差 .",1.0979772646635409e-10,0

不错，在普宁寺旁边，有特色。服务员态度特别好，很亲切，都穿藏服。,不错 ， 普宁寺 旁边 ， 特色 。 服务员 态度 特别 好 ， 亲切 ， 穿 藏服 。,0.9751382761488936,1

预订海景花园大酒店已经没有房间了，于是换到了这里，看到外表有些疑惑，因为酒店并不大，入住时却得到了出乎意料的惊喜：前台服务很温馨，cheakin的速度很快，门童很热情，主动给你提行李，如果开车去会记住为你留好车位；房间面积不小，装饰得很精致，还有果盘赠送；环境稍差些，看不到海景，但是从酒店出来，穿过马路一直向海走，再过一条马路就到了著名的韩式烧烤---李朝木炭烧烤了，另外一点不足就是旁边有一家KTV，晚上能感觉到有些吵。总体感觉很不错，可以给4星。,预订 海景 花园 大酒店 已经 没有 房间 ， 于是 换到 这里 ， 看到 外表 有些 疑惑 ， 因为 酒店 并不大 ， 入住 时 却 得到 出乎意料 惊喜 ： 前台 服务 温馨 ， cheakin 速度 很快 ， 门童 热情 ， 主动 给 你 提 行李 ， 如果 开车 去 会 记住 为 你 留好 车位 ； 房间 面积 不小 ， 装饰 得 精致 ， 还有 果盘 赠送 ； 环境 稍差 些 ， 看不到 海景 ， 但是 从 酒店 出来 ， 穿过 马路 一直 向海 走 ， 再 过 一条 马路 到 著名 韩式 烧烤 --- 李朝 木炭 烧烤 ， 另外 一点 不足 就是 旁边 一家 KTV ， 晚上 能 感觉 到 有些 吵 。 总体 感觉 不错 ， 可以 给 4 星 。,0.9965434706983288,1

很一般的五星级酒店，周边的环境感觉很乱,一般 五星级 酒店 ， 周边 环境 感觉 乱,0.4605964088646638,0

宝晖酒店离机场比较近，不用到罗湖一带找酒店了酒店大堂和餐厅等装璜都挺漂亮的，很有特色的古雅，服务人员也很不错，热情周到，这点没什么挑剔但是，内在客房设施不大满意我住的行政房间，居然在写字台附近没有插座！这样有宽带又有什么用？只好把电视机柜挪开才能给电脑充电，真不方便其他设备也比较缺，比如小冰箱、箱包架（后来发现藏在衣柜里了）、面巾纸盒、咖啡...还有一点似乎不够人性化，离开房间拔掉房卡，房间是马上断电的立刻漆黑的一团，如果是CheckOut拖着大包小包更是不方便。很多酒店是拔掉房卡后有10秒到甚至一分钟的延迟才断电，建议借鉴一下,宝晖 酒店 离 机场 比较 近 ， 不用 到 罗湖 一带 找 酒店 酒店 大堂 餐厅 等 装璜 挺 漂亮 ， 特色 古雅 ， 服务 人员 不错 ， 热情周到 ， 这点 没什么 挑剔 但是 ， 内在 客房 设施 大 满意 住 行政 房间 ， 居然 写字台 附近 没有 插座 ！ 这样 宽带 又 什么 用 ？ 只好 把 电视机 柜 挪开 才能 给 电脑 充电 ， 真 方便 其他 设备 比较 缺 ， 比如 小 冰箱 、 箱包 架 （ 后来 发现 藏 衣柜 里 ） 、 面巾纸 盒 、 咖啡 ... 还有 一点 似乎 不够 人性化 ， 离开 房间 拔掉 房卡 ， 房间 马上 断电 立刻 漆黑 一团 ， 如果 CheckOut 拖 着 大 包小包 更是 方便 。 很多 酒店 拔掉 房卡 后 10 秒 到 甚至 一分钟 延迟 才 断电 ， 建议 借鉴 一下,3.7549052755991847e-10,0

连续入住好几次了。将来可能没有机会再到湖南这家店却是还可以,连续 入住 好 几次 。 将来 可能 没有 机会 再 到 湖南 这家 店 却是 还 可以,0.8534592292255302,1

"很不错的一个酒店,床很大,很舒服.酒店员工的服务态度很亲切.","不错 一个 酒店 , 床 很大 , 舒服 . 酒店 员工 服务态度 亲切 .",0.9072162816102792,1

客房环境级差，房间里时常可以看见“小强”，新铺的床单上由占整个面积3分之2的不知道是什么液体留下的印记，还夹杂着3、4个黑鞋印，恶心之极。卫生间所有的下水都很慢，在浴缸里冲澡每5、6分钟得听一小会儿，要不就水淹七军了，毛巾都是灰色的，实在是差劲。建议想入住德请慎重考虑。,客房 环境 级差 ， 房间 里 时常 可以 看见 “ 小强 ” ， 新铺 床单 上 由 占 整个 面积 3 分 之 2 知道 什么 液体 留下 印记 ， 还 夹杂着 3 、 4 个 黑 鞋印 ， 恶心 之极 。 卫生间 所有 下水 很慢 ， 浴缸 里 冲澡 每 5 、 6 分钟 得 听 一小 会儿 ， 要 水淹七军 ， 毛巾 灰色 ， 实在 差劲 。 建议 想 入住 德请 慎重考虑 。,1.2041478925084448e-12,0

## Sample for annotation: samples\_for\_annotation\_ecommerce.csv

review,tokens\_join,sentiment\_score,sentiment\_label

洗头水不错，洗后头发柔软顺滑，618便宜很多………,洗头水 不错 ， 洗后 头发 柔软 顺滑 ， 618 便宜 很多 … … …,0.0765316588374701,0

真的让我很失望,真的 让 失望,0.1562016488867184,0

和图上完全不像,图上 完全 不像,0.4461576432852818,0

感觉还可以，挺流畅的，电量暂时还没彻底放开用过，暂时不知道怎么样,感觉 还 可以 ， 挺 流畅 ， 电量 暂时 还 没 彻底 放开 用过 ， 暂时 知道 怎么样,0.9608055016305412,1

昨天拍的，今天上午就到了，物流超级快，收到后很惊喜，物超所值，快递小哥每次都送到单位，非常感谢您～,昨天 拍 ， 今天上午 到 ， 物流 超级 快 ， 收到 后 惊喜 ， 物超所值 ， 快递 小哥 每次 送到 单位 ， 非常 感谢您 ～,0.9967642490885128,1

第一次在京东买水果！保鲜方式挺好的，水果也不错，只是没有想象中的大。,第一次 京东 买 水果 ！ 保鲜 方式 挺 好 ， 水果 不错 ， 只是 没有 想象 中 大 。,0.9976986316703472,1

货单打开一看吓人吖！一件30码、一件32码！原来是下错单了啦，大两码了这样一来让我怎么穿的好,货单 打开 一看 吓人 吖 ！ 一件 30 码 、 一件 32 码 ！ 原来 下 错单 啦 ， 大 两码 这样一来 让 怎么 穿 好,0.0520425535957602,0

京东配送服务先来说，上午的单子，下午送达，很赞的。是正品，味道和之前用的一样，赞。性价比也很高。好评,京东 配送 服务 先 来说 ， 上午 单子 ， 下午 送达 ， 很赞 。 正品 ， 味道 之前 用 一样 ， 赞 。 性价比 高 。 好评,0.999906165700128,1

京东买比超市便宜，还送货上门！关键是很甜！,京东 买 比 超市 便宜 ， 还 送货上门 ！ 关键 很甜 ！,0.984048352184418,1

无宽带成为最大问题，周边无工商银行成为第二问题。补充点评 2007年1月13日 ： 无可吃的早餐吃成为不可原谅问题。,无 宽带 成为 最大 问题 ， 周边 无 工商银行 成为 第二 问题 。 补充 点评 2007 年 1 月 13 日 ： 无可 吃 早餐 吃 成为 不可 原谅 问题 。,0.0001438680605826,0

## Sample for annotation: samples\_for\_annotation\_waimai.csv

review,tokens\_join,sentiment\_score,sentiment\_label

每次都是，打电话说餐到了，结果下楼等了很久才到！送餐员就是怕等，让我等很久！！！,每次 ， 打电话 说餐 到 ， 结果 下楼 等 久 才 到 ！ 送 餐员 就是 怕 等 ， 让 等 久 ！ ！ ！,0.051293279659258,0

一个多小时才送到，超级慢，到了东西也凉了，厚厚一层油挂着，根本就没法吃了！！！送餐的人说是不认识路，找了很久，怎么着？地址写得好好的，自己业务不熟路都不认识就成了来晚了的理由了？没天理了，太过分！！再也不会用了！！！,一个多 小时 才 送到 ， 超级 慢 ， 到 东西 凉 ， 厚厚 一层 油挂 着 ， 根本 没法 吃 ！ ！ ！ 送餐 说 认识 路 ， 找 久 ， 怎么着 ？ 地址 写得 好好 ， 自己 业务 熟路 认识 成 来 晚 理由 ？ 没 天理 ， 太过分 ！ ！ 再也不会 用 ！ ！ ！,3.956879972677463e-06,0

太难吃了。,太 难吃 。,0.2733303707351114,0

有点甜...其他都还好。,有点 甜 ... 其他 还好 。,0.8214667880797474,1

水煮肉味道非常好，菜量足，豆腐鲜嫩可口，方便，服务良好。,水 煮肉 味道 非常 好 ， 菜量 足 ， 豆腐 鲜嫩 可口 ， 方便 ， 服务 良好 。,0.9413104459239352,1

"太慢了,说好45分钟,结果一个多小时","太慢 , 说好 45 分钟 , 结果 一个多 小时",0.0491080901182352,0

丽华饭量小了点,丽华 饭量 小 点,0.3881389894600598,0

"口味很不错,但是卤蛋真的很坑,建议别点","口味 不错 , 但是 卤蛋 真的 很坑 , 建议 别点",0.5366758138582177,1

快递小哥很给力！冒雨前来，态度很好！赞！！！！,快递 小哥 给 力 ！ 冒雨 前来 ， 态度 好 ！ 赞 ！ ！ ！ ！,0.7081745744028077,1

只送来了卷饼但是没送粥，打电话一直占线，最起码打了不下十次了，总算通了又让我打另一个电话，又继续不是占线就没人接，终于打通了才把粥给我补上送来，据我下单已经过去快两个小时了,只 送来 卷 饼 但是 没送 粥 ， 打电话 一直 占线 ， 最起码 打 不下 十次 ， 总算 通了 又 让 打 另 一个 电话 ， 又 继续 不是 占线 没人接 ， 终于 打通 才 把 粥 给 补上 送来 ， 据 下单 已经 过去 快 两个 小时,6.25158414147009e-07,0