

# 机器学习大作业报告

SJTU

---

Name: 余俊洁  
Student Number: 523030910244  
Research Project Title: 机器学习大作业

---

## 目录

1 大作业概论	2
2 数据与任务描述	2
2.1 数据来源与字段简介	2
2.2 预测任务	2
3 数据预处理	2
4 特征学习	2
4.1 PCA (回归任务)	3
4.2 LDA (分类任务)	3
4.3 SelectKBest (回归与分类任务)	3
5 实验数据集合集	3
6 实验设计	4
6.1 滚动预测策略	4
6.2 超参数网格	4
6.2.1 回归模型	4
6.2.2 分类模型	5
6.3 评估指标	5
6.4 可视化	5
7 实验结果与分析	6
7.1 回归任务结果与分析	6
7.2 分类任务结果与分析	9
附录：实验结果	13
A 回归任务结果	13
A.1 线性模型	13
A.2 Lasso 回归模型	15
A.3 Ridge 回归模型	18
A.4 随机森林模型	20
A.5 SVR	22

<b>A2分类任务结果</b>	<b>24</b>
A2.1 KNN . . . . .	24
A2.2 Logistic . . . . .	28
A2.3 LDA . . . . .	32
A2.4 贝叶斯 . . . . .	36
A2.5 随机森林 . . . . .	39
A2.6 SVM . . . . .	42

# 1 大作业概论

本次机器学习大作业针对 2015-2019 年旅客出行数据，完成了对于停留天数（回归任务）和旅游目的（分类任务）两项预测任务。首先我对 5 年的原始数据进行了合并，清洗与归一化处理等工作；随后，我使用特征学习技术，分别对数据进行了针对回归任务的主成分分析（PCA）降维，和针对分类任务的线性判别分析（LDA）降维。在时间序列背景下，按照要求，使用 `sklearn.model_selection.TimeSeriesSplit(n_splits = 5)` 对时间序列数据进行多折划分进行模型评估与网格化超参数搜索。其中，我使用的回归模型包括普通线性模型、L1（Lasso）、L2（Ridge）、随机森林和 SVR；分类模型包括 KNN、Logistic 回归、LDA、贝叶斯、SVM 和随机森林。最后，对比分析了在不同的数据处理方式下，不同模型在多维指标上的表现，并讨论特征学习与正则化对模型稳定性的影响。

## 2 数据与任务描述

### 2.1 数据来源与字段简介

数据包含 2015-2019 年共 (72370, 382) 组数据（将所有数据集合并之后的`.shape`）。主要特征组如下：

- **类别变量：**包括游客的国籍、性别、旅行类型、最满意/最向往的地点、满意度水平等核心属性
- **数值变量：**包括游客的年龄、消费情况、旅游频次及多个旅游项目的满意度打分等
- **二值变量：**涉及游客在旅途中涉及的住宿、交通、活动、目的地、旅游动机及景点访问等行为特征

### 2.2 预测任务

**任务 1：** 回归——预测连续型目标变量：旅客停留天数 (*Number of nights in CITY*)

**任务 2：** 分类——预测类别型目标变量：旅客旅游目的 (*Purpose of visit to CITY*)

## 3 数据预处理

1. **加载并合并数据：** 将 2015 年到 2019 年的所有数据合并并对其按照 ‘Survey date’ 顺序排序。
2. **定义特征类型：** 分析数据，将数据分为类别变量，数值变量，二值变量三类。
3. **缺失值处理：** 连续变量采用平均值填补，类别变量和二值变量采用众数填补。
4. **类别编码：** One-Hot (保留稀疏性)。
5. **训练数据构建：** 构建用于回归和分类任务的训练数据，根据目标变量是否缺失来筛选样本，并构建特征 (X) 和目标变量 (y)。
6. **标准化：** 对特征数据进行 Z-score 标准化，以提高模型训练的效果和稳定性。

## 4 特征学习

降维或特征学习的目标是从原始特征  $\mathbf{X}_{\text{orig}} \in \mathbb{R}^{n \times p}$  中提取更具代表性的信息，提升模型性能与泛化能力。在实验中，我采用了三种主流特征学习方法：PCA、LDA 以及基于统计评分的 SelectKBest，分别用于回归和分类任务。

## 4.1 PCA (回归任务)

PCA 通过线性变换将高维特征空间投影到低维空间，并尽可能保留原始数据的方差信息：

$$\mathbf{Z}_{\text{PCA}} = \mathbf{X}_{\text{orig}} \mathbf{W}_k, \quad \mathbf{W}_k = \arg \max_{\mathbf{W}^\top \mathbf{W} = I} \text{Var}(\mathbf{X}\mathbf{W})$$

对于回归任务的数据，我采用了 PCA 降维，在多次尝试之后，最终选定降维后的特征维度为 20。

## 4.2 LDA (分类任务)

LDA 利用类别标签信息，在类别间最大化区分度的同时最小化类别内差异。目标函数如下：

$$\mathbf{W}_{\text{LDA}} = \arg \max_{\mathbf{W}} \frac{\det(\mathbf{W}^\top S_B \mathbf{W})}{\det(\mathbf{W}^\top S_W \mathbf{W})}$$

其中  $S_B$  和  $S_W$  分别为类间散度矩阵和类内散度矩阵。对于分类任务的数据，我采用了 LDA 降维，由于 LDA 输出维度上限为类别数 -1，本任务共有 4 类，因此输出维度为 3。

## 4.3 SelectKBest (回归与分类任务)

SelectKBest 是一种基于统计检验的特征选择方法，通过评分函数评估每个特征与目标变量的相关性，选择得分最高的前  $k$  个特征。本文采用了两种评分函数：

- 回归任务使用 **F-statistic** (`f_regression`)，评估每个特征与目标变量之间的线性相关性。
- 分类任务使用 **ANOVA F 值** (`f_classif`)，评估每个特征对类别标签的区分能力。

为比较不同的  $k$  的选择下的效果，我在实验中设置了两个  $k$  值： $k = 10$  与  $k = 50$ ，并分别构建了回归和分类任务下的特征子集  $\mathbf{X}_{\text{KBest}}$ 。

**注：**

当我在用上述数据训练模型时发现，会有一些极端值影响模型的效果，因而我们需要对那些极端值进行筛选，我尝试过直接对原始数据集进行极端值过滤，但是原始数据拥有 382 个维度，筛选过程消耗了大量不必要的算力，更重要的是，很多数据可能因为一些在那些不那么重要的数据上的离群值而被排除，进而损失了大量的有效的数据。

因此，我考虑在我得到的 SelectKBest 的数据集上进行按 Z-score 的异常值清洗，这样能够更为精准地去除那些我们不需要的数据且复杂度更低。同时，在实验中发现，对于  $k=10$  或  $50$ ，两者的效果接近， $k=10$  略胜一筹且复杂度更低，故实验中选取的极端值清洗后的数据集是基于  $k=10$  的特征提取数据集的。

## 5 实验数据集合集

经过数据预处理和特征学习过程，选取得到以下进行实验的数据集：

- 回归任务：
  - $\mathbf{X}^{(\text{reg-init})} \in \mathbb{R}^{n \times p}$ , 原始数据;
  - $\mathbf{X}^{(\text{reg-sca})} \in \mathbb{R}^{n \times p}$ , 标准化后的数据;
  - $\mathbf{X}^{(\text{reg-pca})} \in \mathbb{R}^{n \times 20}$ , PCA 后的数据 (经过标准化);
  - $\mathbf{X}_{10}^{(\text{reg-selected})} \in \mathbb{R}^{n \times 10}$ , 10 个主要特征提取后的数据;
  - $\mathbf{X}_{50}^{(\text{reg-selected})} \in \mathbb{R}^{n \times 50}$ , 50 个主要特征提取后的数据;

–  $\mathbf{X}_{10}^{(\text{reg-selected-cleaned})} \in \mathbb{R}^{n' \times 10}$ , Z-score 清洗后的 10 特征提取数据。

- 分类任务:

- $\mathbf{X}^{(\text{cls-init})} \in \mathbb{R}^{m \times p}$ , 原始数据;
- $\mathbf{X}^{(\text{cls-sca})} \in \mathbb{R}^{m \times p}$ , 标准化后的数据;
- $\mathbf{X}^{(\text{cls-lda})} \in \mathbb{R}^{m \times 3}$ , LDA 后的数据 (经过标准化);
- $\mathbf{X}_{10}^{(\text{cls-selected})} \in \mathbb{R}^{m \times 10}$ , 10 个主要特征提取后的数据;
- $\mathbf{X}_{50}^{(\text{cls-selected})} \in \mathbb{R}^{m \times 50}$ , 50 个主要特征提取后的数据;
- $\mathbf{X}_{10}^{(\text{cls-selected-cleaned})} \in \mathbb{R}^{m' \times 10}$ , Z-score 清洗后的 10 特征提取数据。

其中  $n, m$  为原始样本数,  $n', m'$  为极端值清洗后样本数。

## 6 实验设计

### 6.1 滚动预测策略

滚动预测策略的目的是将数据划分为多个训练集和测试集, 其中每次的训练集包含了当前时间点之前的数据, 而测试集则包含了之后的数据, 这样能够有效模拟实际的预测场景, 即利用过去的数据预测未来的趋势。

在实验中, 按照要求, 我使用 `sklearn.model_selection.TimeSeriesSplit(n_splits = 5)` 来进行时间序列的交叉验证, 根据第一列数据变量 Survey date 对数据进行切片。

---

#### Algorithm 1 五折滚动预测

---

- 1: 设时间序列样本按日期升序排列
  - 2: **for**  $i = 1$  **to** 5 **do**
  - 3:    使用前  $i$  个时间窗口作为训练集, 下一窗口作为验证集
  - 4:    进行网格搜索 & 模型训练
  - 5:    记录折内最佳模型及性能
  - 6: **end for**
  - 7: 计算网格搜索得到的最佳参数对应模型所有折指标的平均值
- 

### 6.2 超参数网格

#### 6.2.1 回归模型

表 1: 回归模型与超参数网格

模型	网格超参数
线性模型	—
Lasso(L1 正则)	$\alpha \in \{10^{-2}, 10^{-1}, 10^0, 10^1\}$
Ridge(L2 正则)	$\alpha \in \{10^{-2}, 10^{-1}, 10^0, 10^1\}$
SVR	$C \in \{0.1, 1, 10\}$ , $\epsilon \in \{10^{-2}, 10^{-1}\}$ , kernel $\in \{\text{'linear'}, \text{'rbf'}\}$
RandomForest	$n\_estimators \in \{100, 200\}$ , $\text{max\_depth} \in \{5, 10, \text{None}\}$ , $\text{min\_samples\_split} \in \{2, 5\}$

### 6.2.2 分类模型

表 2: 分类模型与超参数网格

模型	网格超参数
KNN	$k \in \{3, 5, 7, 9\}$ , weights $\in \{\text{'uniform'}, \text{'distance'}\}$ , $p \in \{1, 2\}$
Logistic	$C \in \{0.01, 0.1, 1, 10\}$ , penalty $= \{l1, l2\}$
LDA (Classifier)	solver $\in \{\text{'svd'}, \text{'lsqr'}, \text{'eigen'}\}$ , shrinkage $\in \{\text{auto}, 0.1, 0.2, 0.3\}$
GaussianNB(贝叶斯)	var_smoothing $\in [10^{-9}, 10^{-8}, 10^{-7}]$
SVM	$C \in \{0.1, 1, 10\}$ , kernel $\in \{\text{'linear'}, \text{'rbf'}\}$ , $\gamma \in \{\text{'scale'}, \text{'auto'}\}$
RandomForest	n_estimators $\in \{100, 200\}$ , max_depth $\in \{5, 10, \text{None}\}$ , min_samples_split $\in \{2, 5\}$

## 6.3 评估指标

回归

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2, \quad \text{MAE} = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|, \quad R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2} \quad (1)$$

选择这些指标，能较好地反映模型对于大误差的敏感性 (MSE)、整体稳健性 (MAE) 和拟合能力 (R2)。

分类

分别采用宏平均与加权平均的 Accuracy / Precision / Recall / F1，以及 OvR AUC，以全面评估模型的整体表现与各类识别能力。

同时实验结果的详细信息中给出了 macro 和 weighted 两种取平均的方式，但在下一节的整体汇总中仅展示了 weighted 的结果，是因为按使用样本量加权的方式，能更真实地反映总体性能，避免被少数类扭曲，也更符合实际的要求。

## 6.4 可视化

回归

- 数据实际值与预测值对比图
- 不同模型在不同的数据集下的评价指标条形图
- 不同模型在不同数据集下得到的不同评估指标的热力图
- R2 雷达图
- MSE,MAE,R2 气泡图
- MSE,MAE,R2 平行坐标图

分类

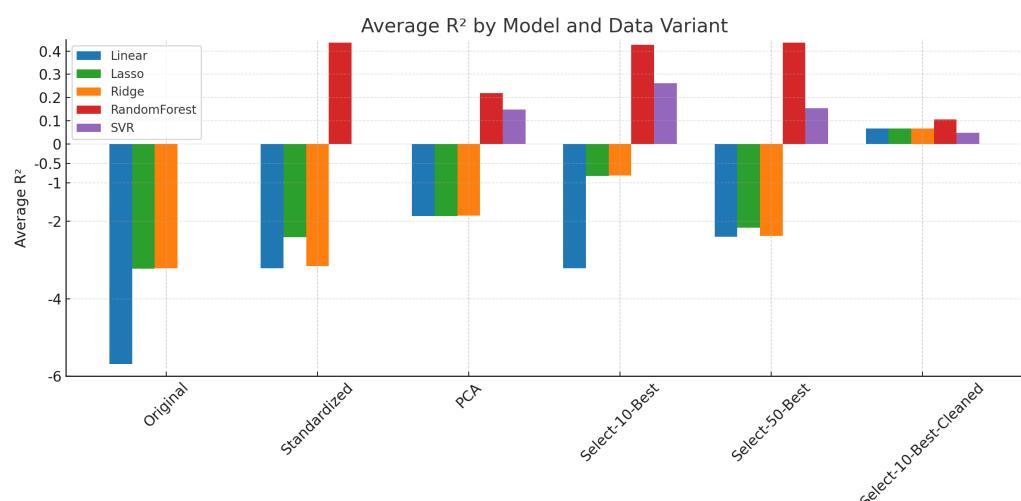
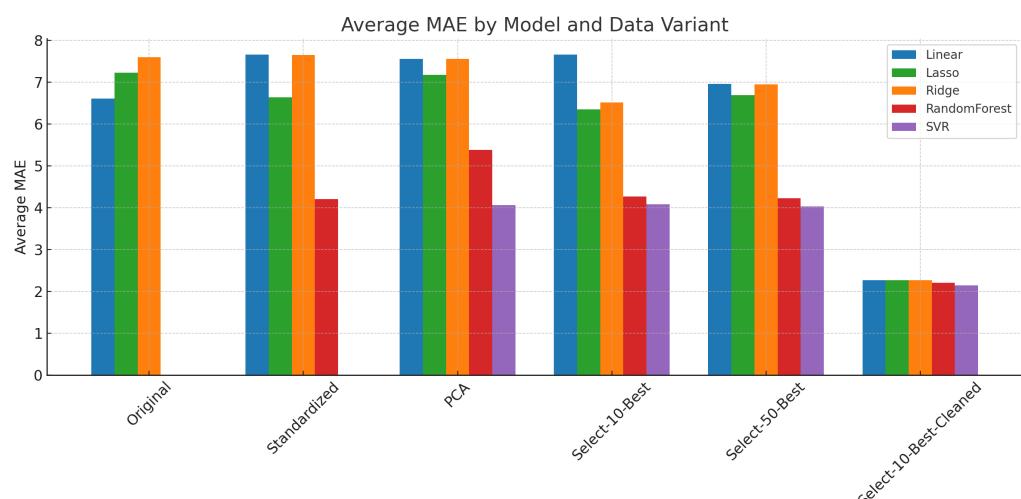
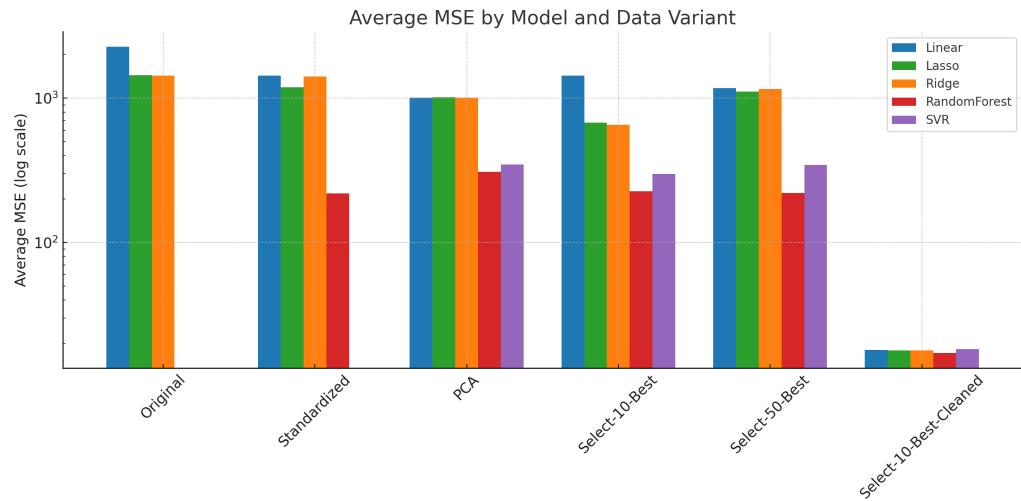
- 不同模型在不同的数据集下的各个评价指标条形图
- Accuracy 雷达图
- 混淆矩阵 (Confusion Matrix)
- 不同类别的 ROC 曲线

## 7 实验结果与分析

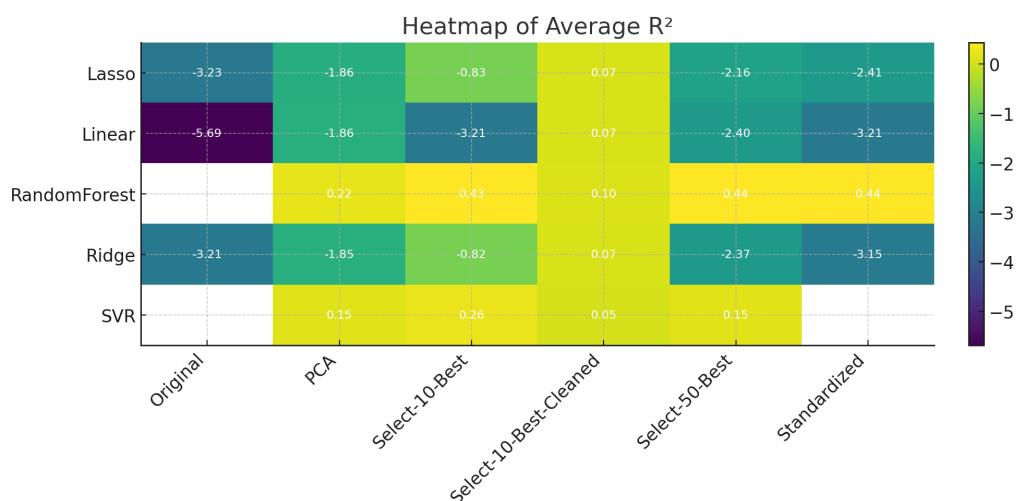
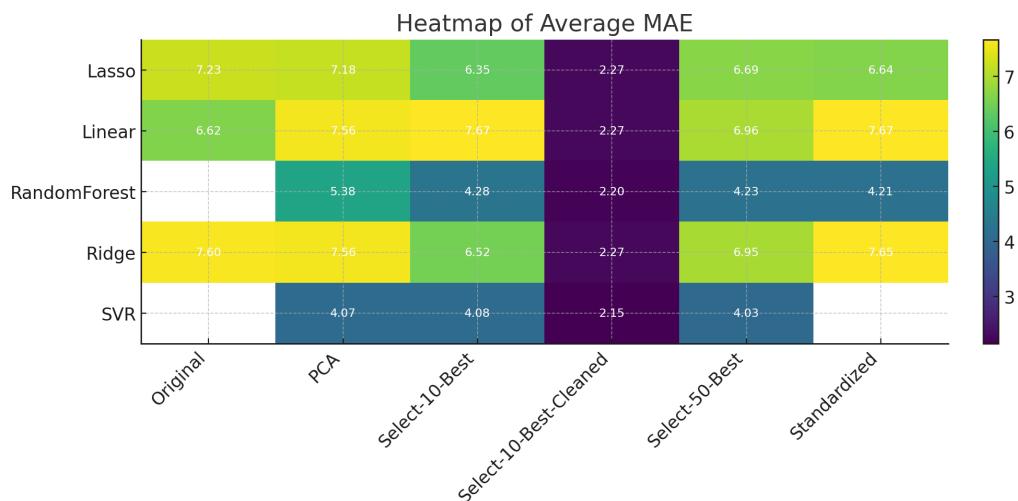
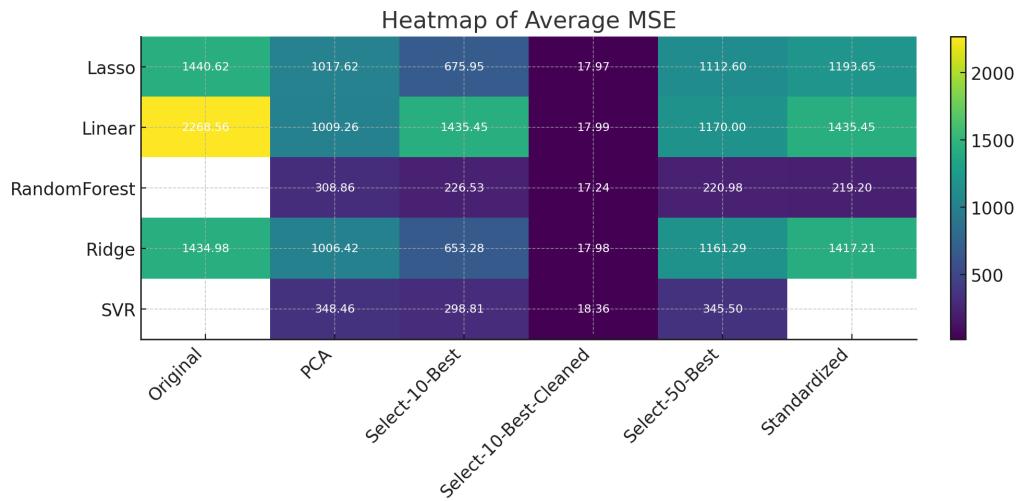
### 7.1 回归任务结果与分析

结果 **注：**不同模型在不同数据集下的具体的数据请参考附录 A.

将所有的实验数据进行归纳，得到以下三种评价指标的可视化呈现：

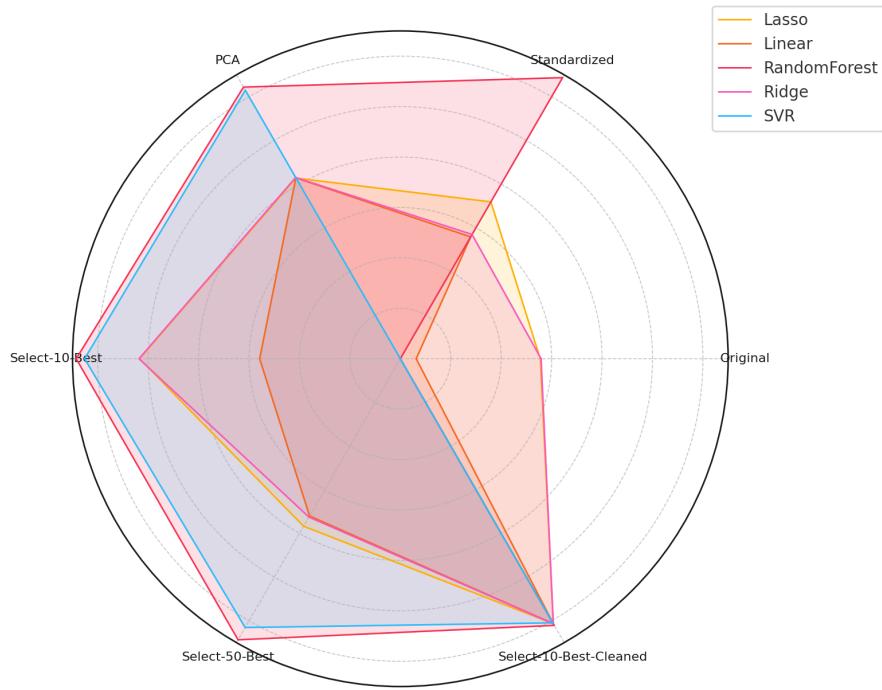


更为直观地，我绘制了不同模型在不同数据集下得到的不同评估指标的热力图：

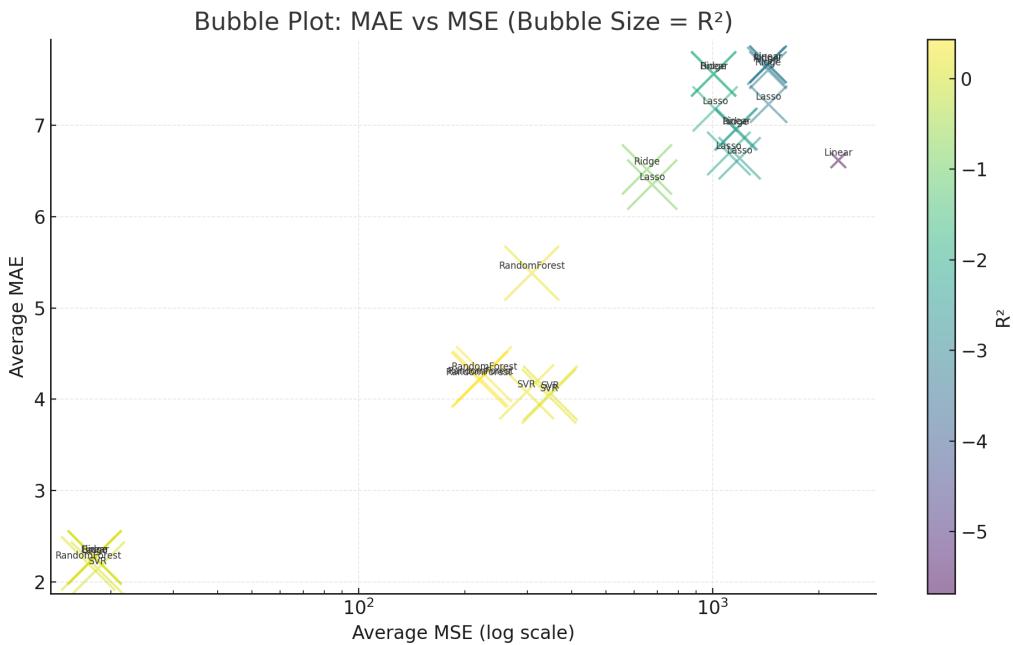


同时，通过绘制 R<sup>2</sup> 的雷达图，我们也能更为清晰地比较各模型的性能：

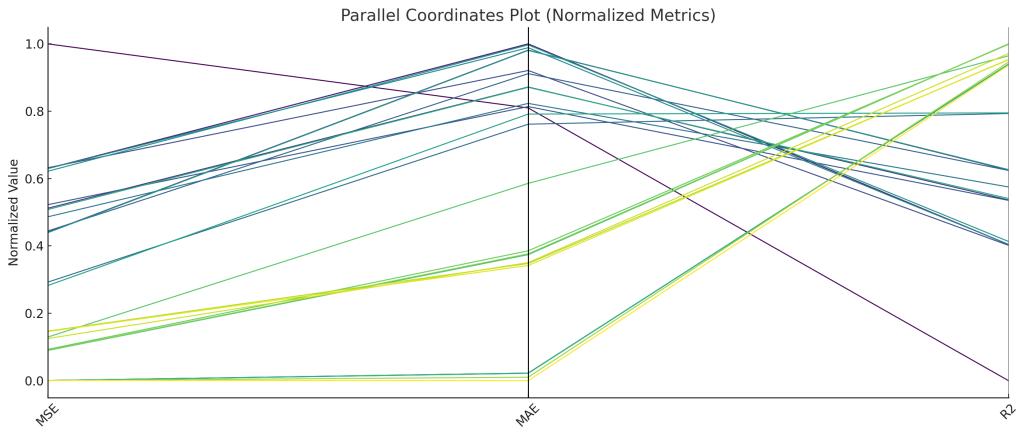
Average R<sup>2</sup> Radar Chart



同时把 MSE,MAE 和 R<sup>2</sup> (气泡大小与颜色) 映射到一张图上, 得到三者的气泡图, 更为便捷地观察模型 - 数据组合误差大小、拟合好坏:



为了直观揭示多指标之间的权衡与优势走向, 我把各组合在 MSE、MAE、R<sup>2</sup> 三条坐标轴上的归一化数值连成折线, 绘制得到平行坐标图:



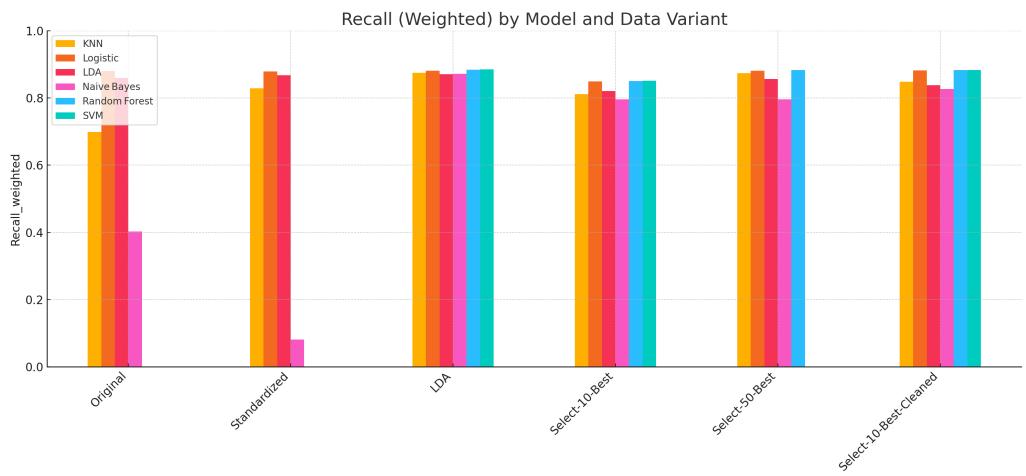
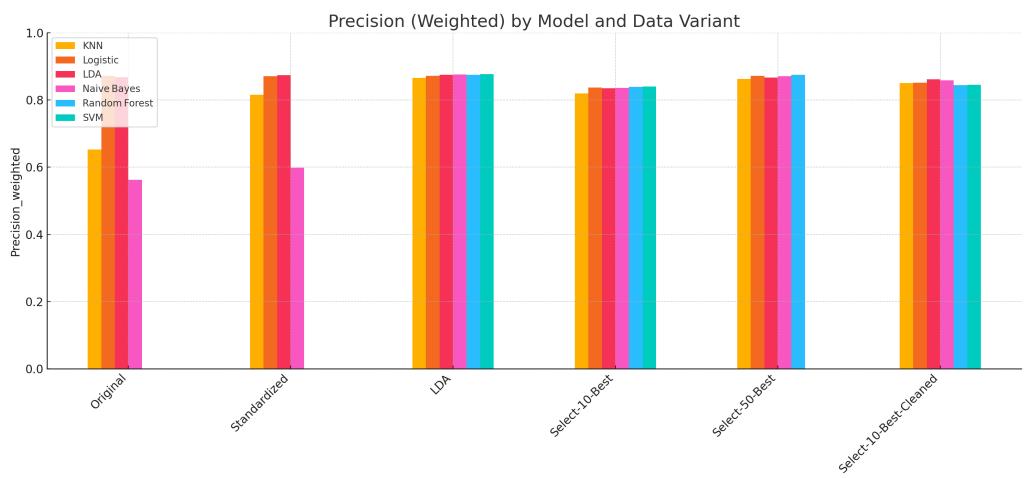
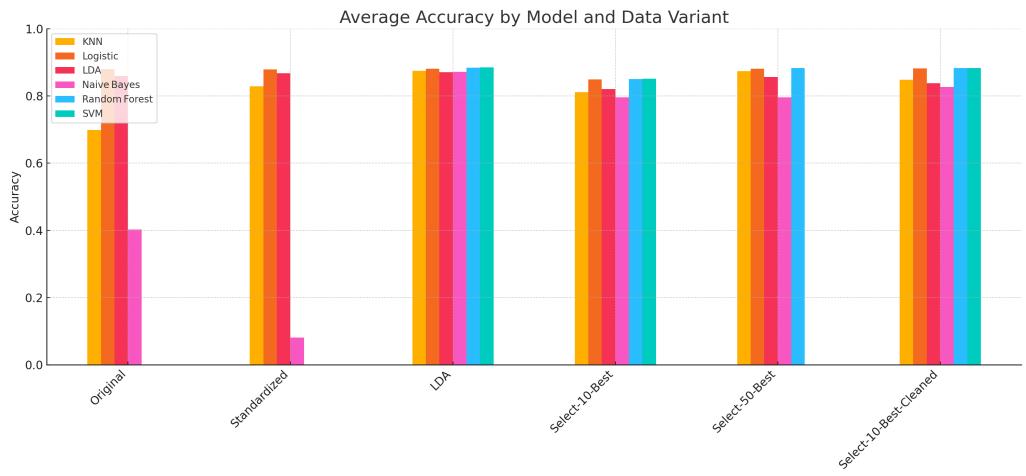
## 分析

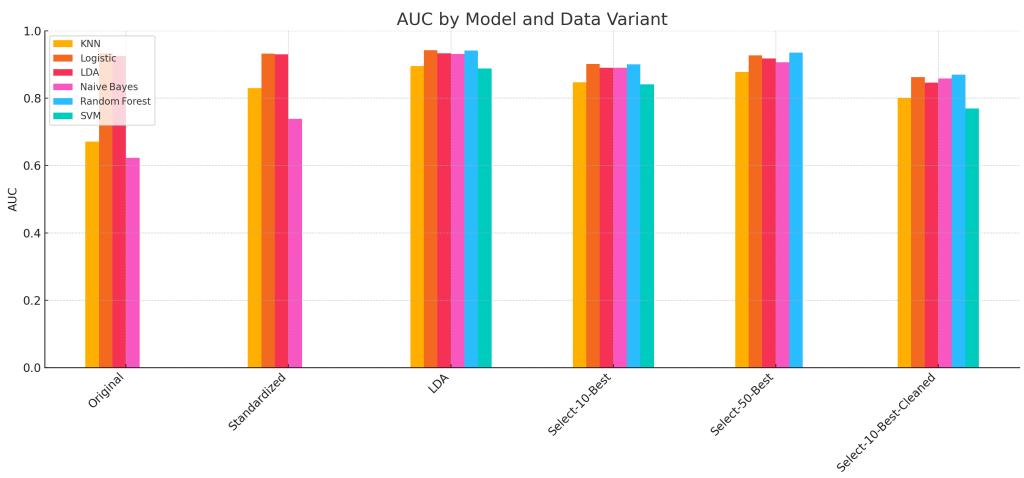
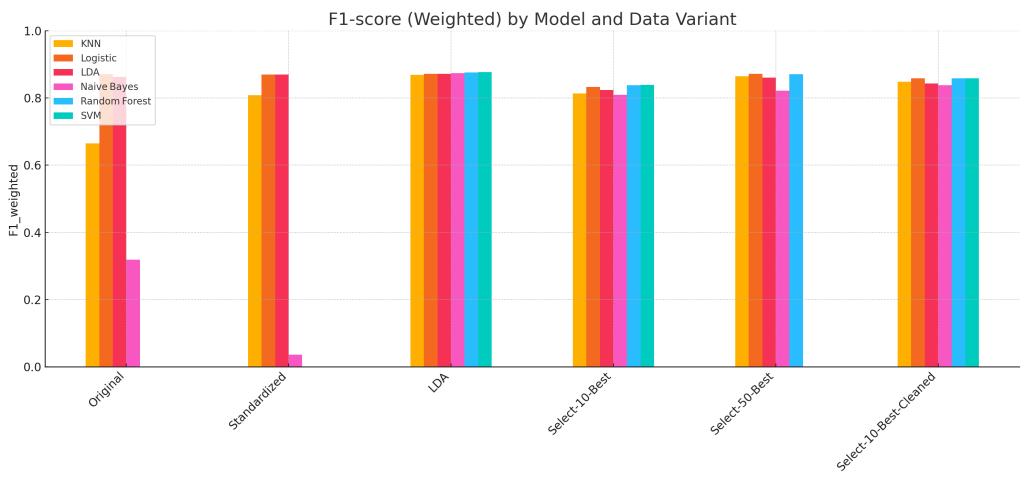
- 模型的适应性与复杂性
  - 线性模型与正则化 (Lasso、Ridge): 虽然这些模型整体的效果很难比得上像随机森林一样的复杂模型，但其的复杂度较低。在数据特征较简单、线性相关性强的情况下表现尚可，然而，在面对高维复杂数据时（如实验中的未经过 PCA 或特征选择的数据集），这些模型的效果就不太令人满意了。特别是当我在进行 Lasso 的超参数调节时， $\alpha$  值过小时，出现了过拟合现象。
  - 随机森林回归模型: 其优异表现的根本原因在于其对非线性关系和高维数据的处理能力。其通过多决策树的集成，有效减少了模型对噪声的敏感性。相比其他回归模型，其对特征选择后的数据表现更为稳定，具有较强的泛化能力。
- 特征工程对回归性能的影响
  - PCA 的降维: PCA 通过减少特征空间维度，有助于减轻计算负担，尤其是在处理高维数据时。然而，这种降维方法并没有在所有模型中带来提升，比如随机森林在 PCA 数据下训练的效果反而还不如直接在标准化后数据下训练的效果，可能在 PCA 的过程中丢失一些对回归任务至关重要的信息，而影响了模型的效果。
  - 在进行极端异常值的清理之后，我发现，所有模型的 R2 均为正了，表明了对于极端离群值的去除能够显著地减少很多干扰，而使得模型的训练效果提升。
  - 特征选择 (SelectKBest): 特征选择帮助剔除不重要的特征，从而降低了数据的噪声水平，提升了回归模型的效果，也很大程度地提高了模型的训练速度。
- 超参数与正则化的敏感性
  - 在实验的过程中，我发现 SVR 在不同的 C 和  $\gamma$  值设置下，表现波动较大，可见超参数调优的重要性。

## 7.2 分类任务结果与分析

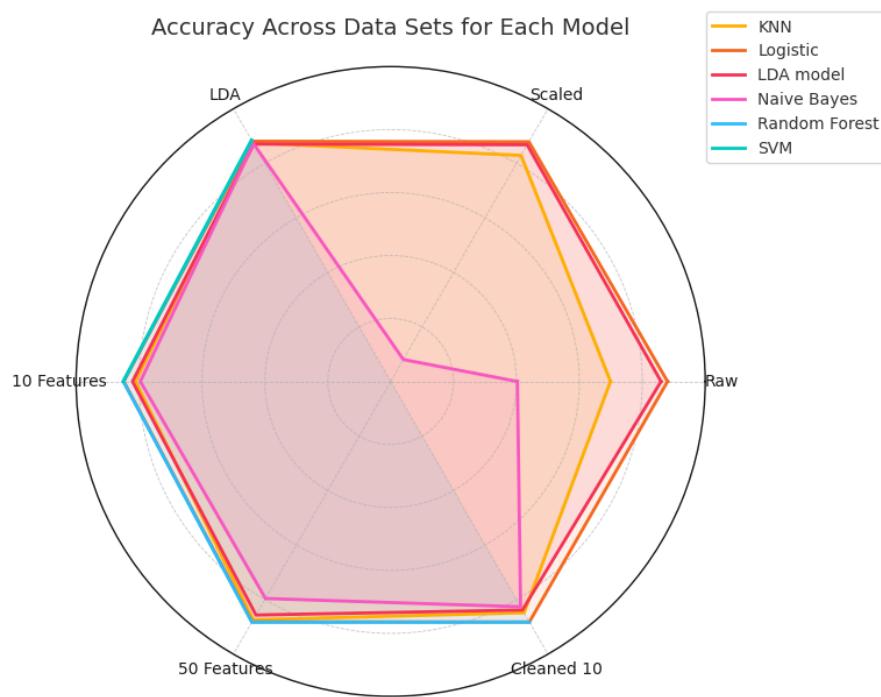
**结果** **注**: 不同模型在不同数据集下的具体的数据与对应的混淆矩阵请参考附录 A.

将所有的实验数据进行归纳，得到以下指标 (Accuracy, Precision(weighted), Recall(weighted), F1-score(weighted), AUC) 的可视化呈现：





为了更直观地看出各个模型在不同数据集下的训练效果，我同样绘制了关于 Accuracy 的雷达图：



## 分析

- 模型的选择与数据集特征
  - KNN 与 Logistic 回归：这些模型在面对标准化后的数据和 LDA 降维后的数据表现较好。KNN 对数据的局部特征非常敏感，因此其在特征提取后能显著提高性能。Logistic 回归也在特征选择后的数据集中表现出较高的精度和较低的过拟合趋势。
  - LDA 的降维效果：从实验结果看来，LDA 降维技术能够显著提高分类任务的效果。与 PCA 不同，LDA 通过利用标签信息进行降维，使得类别间的区分度最大化，从而在多个分类模型中展现了显著的性能提升。
- 贝叶斯分类器的局限性
  - 贝叶斯模型的低效表现：从实验结果来看，其在原始数据和标准化的数据下训练得到的性能较差，其原因可能在于贝叶斯模型假设特征之间是独立的，但在实际的旅游数据中，特征之间往往是相关的，导致贝叶斯假设不成立，性能较差。
- 模型集成与鲁棒性
  - 随机森林通过集成学习的方式，能够有效增强模型的鲁棒性，对特征的敏感性较低。SVM 则在 LDA 降维后的数据集上表现尤为突出，尤其在较高维度的特征空间中，SVM 通过选择适当的核函数能够提高模型的分类边界精度。
- 特征选择与数据清洗的影响
  - 经过实验，我们可以发现清洗数据能够显著提升一些模型的泛化能力，在特征提取后的数据集上，这些分类模型的准确性、召回率和 F1 分数均有所提高。说明通过剔除噪声和无关特征，能够减少模型训练中的冗余信息，从而使模型更加专注于对关键特征的学习。

整体来看：

1. 随机森林在回归任务上，Logistic 和随机森林在分类任务上的综合指标最优。
2. PCA 可显著降低训练成本，且对非线性模型有较强的正向效果，但是对于随机森林等非线性模型的正向效果不佳；LDA 对 KNN 的提升较为明显。
3. 超参数搜索的方式能够方便我们找到合适的模型的训练参数，在实验过程中，很明显能感受到精确调节模型的超参数对于提高回归任务的表现的重要性。
4. 数据的处理方式极大程度上地影响了模型的训练效果，回归实验中，R2 负值的出现跟那些离群值密切相关，从实验的结果来看，通过 Z-score 的方式也没能很好地剔除掉一些异常值，后续可进行优化。

## 附录：实验结果

### A 回归任务结果

#### A.1 线性模型

原始数据：

表 3: 线性模型原始数据训练下五折评估指标

折数	MSE	MAE	R2
Fold 1	336.8344	5.7650	0.1689
Fold 2	9770.1011	10.4300	28.7898
Fold 3	336.3178	5.4946	0.0355
Fold 4	346.0066	5.2283	0.0926
Fold 5	553.5520	6.1636	0.0664
Average	2268.5624	6.6163	-5.6853
Best	336.8344	5.7650	0.1689

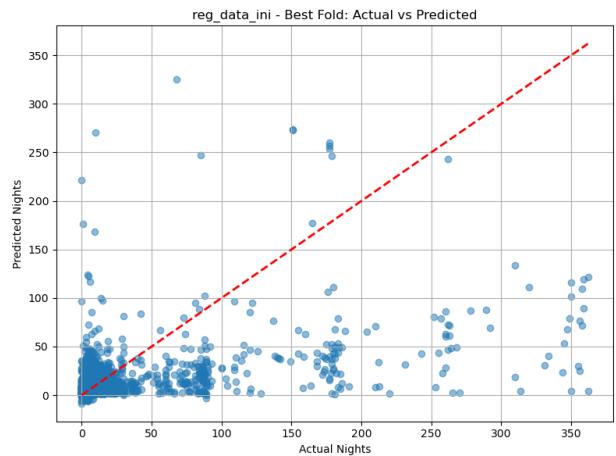


图 1: Actual Nights vs Predicted Nights

标准化数据：

表 4: 线性模型标准化数据训练下五折评估指标

折数	MSE	MAE	R2
Fold 1	263.3822	7.4275	0.3502
Fold 2	5990.6823	11.3312	-17.2660
Fold 3	278.0367	6.4655	0.2026
Fold 4	268.0578	6.3568	0.2971
Fold 5	377.1054	6.7491	0.3640
Average	1435.4529	7.6660	-3.2104
Best	377.1054	6.7491	0.3640

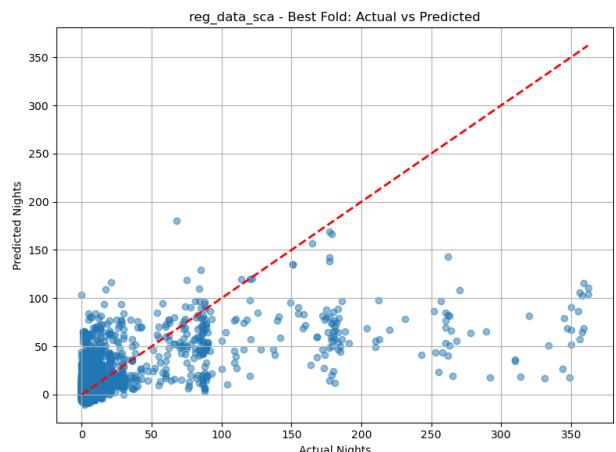


图 2: Actual Nights vs Predicted Nights

## PCA 数据:

表 5: 线性模型 PCA 数据训练下五折评估指标

折数	MSE	MAE	R2
Fold 1	312.1421	7.1290	0.2298
Fold 2	3560.3123	11.3488	-9.8557
Fold 3	307.5704	6.1773	0.1180
Fold 4	341.9268	6.1012	0.1033
Fold 5	524.3376	7.0610	0.1157
Average	1009.2578	7.5635	-1.8578
Best	312.1421	7.1290	0.2298

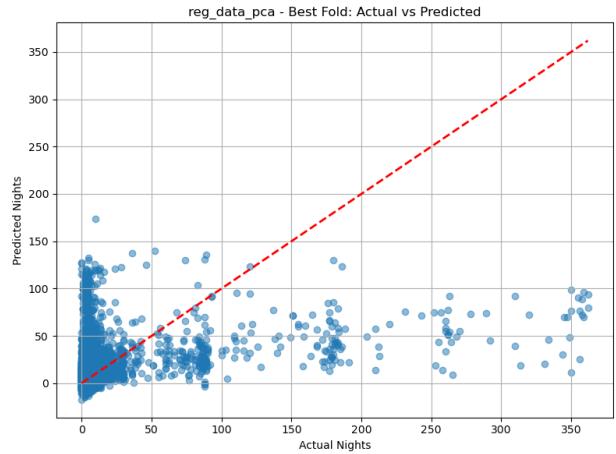


图 3: Actual Nights vs Predicted Nights

## 10 项特征提取数据:

表 6: 线性模型 10 项特征提取数据训练下五折评估指标

折数	MSE	MAE	R2
Fold 1	263.3822	7.4275	0.3502
Fold 2	5990.6823	11.3312	-17.2660
Fold 3	278.0367	6.4655	0.2026
Fold 4	268.0578	6.3568	0.2971
Fold 5	377.1054	6.7491	0.3640
Average	1435.4529	7.6660	-3.2104
Best	377.1054	6.7491	0.3640

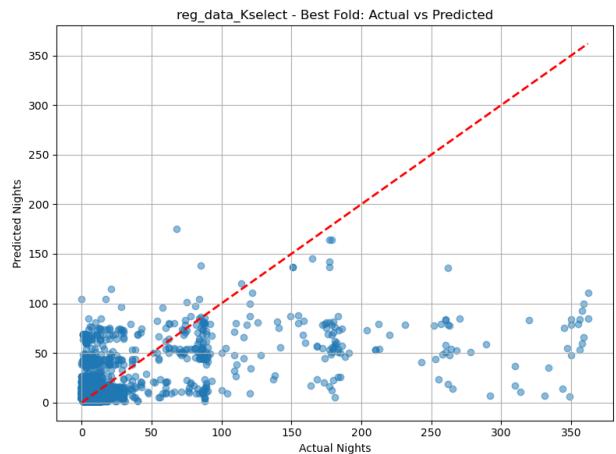


图 4: Actual Nights vs Predicted Nights

## 50 项特征提取数据:

表 7: 线性模型 50 项特征提取数据训练下五折评估指标

折数	MSE	MAE	R2
Fold 1	255.0999	6.6487	0.3706
Fold 2	4664.2164	9.6479	-13.2216
Fold 3	278.7787	5.9347	0.2005
Fold 4	269.4189	6.0050	0.2935
Fold 5	382.4730	6.5599	0.3549
Average	1169.9974	6.9592	-2.4004
Best	255.0999	6.6487	0.3706

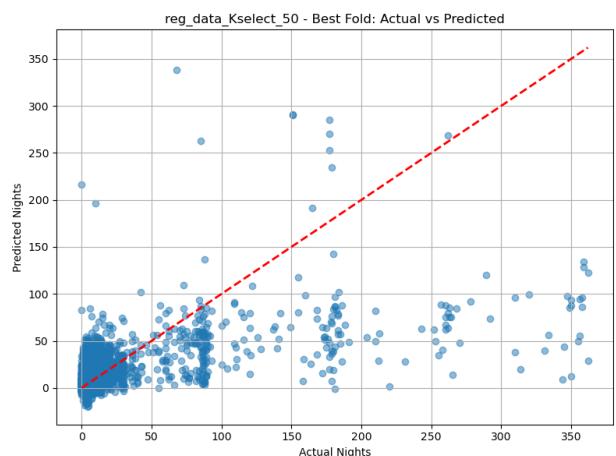


图 5: Actual Nights vs Predicted Nights

清洗后的 10 项特征提取数据:

表 8: 线性模型清洗后的 10 项特征提取数据训练下五折评估指标

折数	MSE	MAE	R2
Fold 1	18.3113	2.2221	0.0742
Fold 2	22.2196	2.4794	0.0018
Fold 3	18.5999	2.3345	0.0931
Fold 4	17.2894	2.1478	0.0835
Fold 5	13.5112	2.1708	0.0798
Average	17.9863	2.2709	0.0665
Best	18.5999	2.3345	0.0931

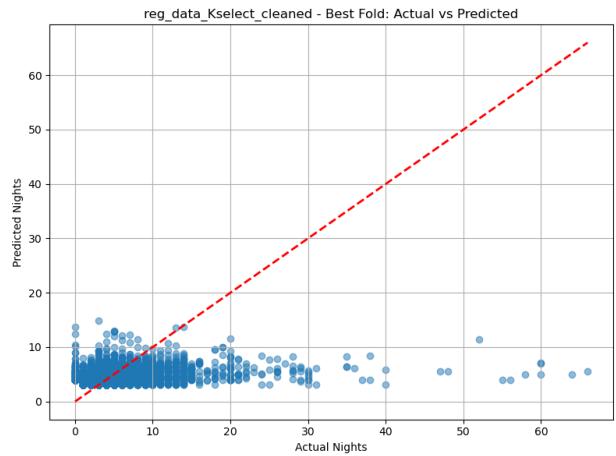


图 6: Actual Nights vs Predicted Nights

## A.2 Lasso 回归模型

原始数据:

表 9: Lasso 回归模型原始数据训练下五折评估指标

折数	MSE	MAE	R2
Fold 1	257.2531	6.3735	0.3653
Fold 2	6008.5252	10.9164	-17.3205
Fold 3	274.7042	6.1136	0.2122
Fold 4	284.7410	6.1634	0.2533
Fold 5	377.8516	6.5836	0.3627
Average	1440.6150	7.2301	-3.2254
Best	257.2531	6.3735	0.3653

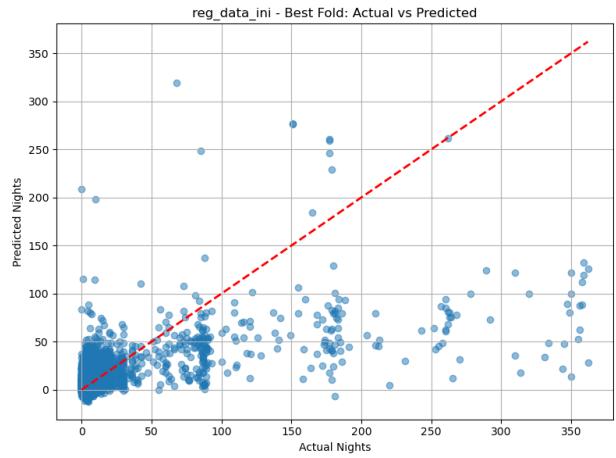


图 7: Actual Nights vs Predicted Nights

标准化数据：

表 10: Lasso 回归模型标准化数据训练下五折评估指标

折数	MSE	MAE	R2
Fold 1	255.2289	6.7233	0.3703
Fold 2	4454.9847	9.6282	-12.5836
Fold 3	348.7290	5.5576	-0.0001
Fold 4	314.9393	4.8643	0.1741
Fold 5	594.3588	6.4180	-0.0024
Average	1193.6481	6.6383	-2.4083
Best	255.2289	6.7233	0.3703

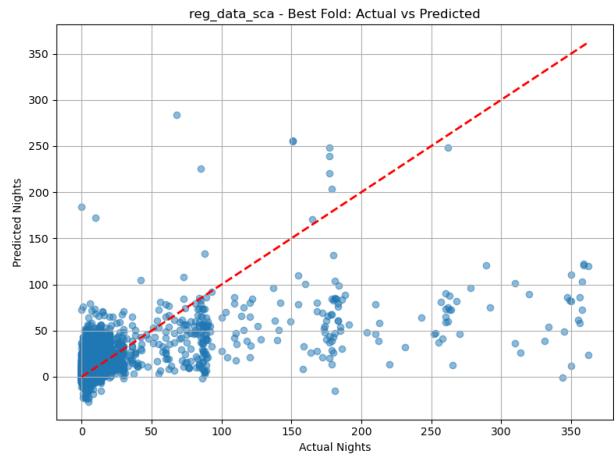


图 8: Actual Nights vs Predicted Nights

PCA 数据：

表 11: Lasso 回归模型 PCA 数据训练下五折评估指标

折数	MSE	MAE	R2
Fold 1	312.6796	6.9356	0.2285
Fold 2	3500.8981	11.2568	-9.6745
Fold 3	343.0862	5.3790	0.0161
Fold 4	341.6980	6.0224	0.1039
Fold 5	589.7213	6.2965	0.0054
Average	1017.6166	7.1781	-1.8641
Best	312.6796	6.9356	0.2285

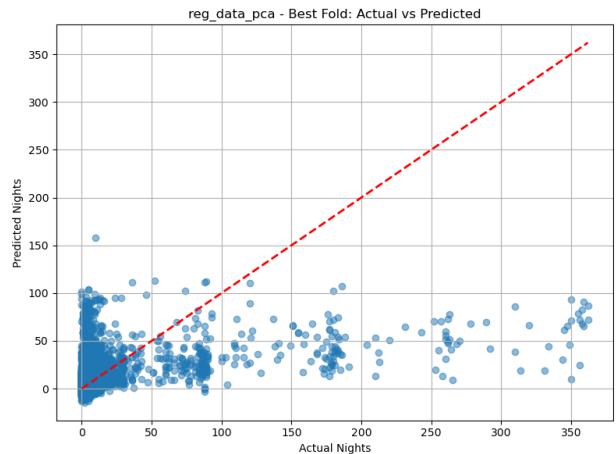


图 9: Actual Nights vs Predicted Nights

10 项特征提取数据：

表 12: Lasso 回归模型 10 项特征提取数据训练下五折评估指标

折数	MSE	MAE	R2
Fold 1	278.8121	6.1926	0.3121
Fold 2	1873.2836	7.9888	-4.7118
Fold 3	348.7290	5.5576	-0.0001
Fold 4	284.5554	5.5936	0.2538
Fold 5	594.3588	6.4180	-0.0024
Average	675.9478	6.3501	-0.8297
Best	278.8121	6.1926	0.3121

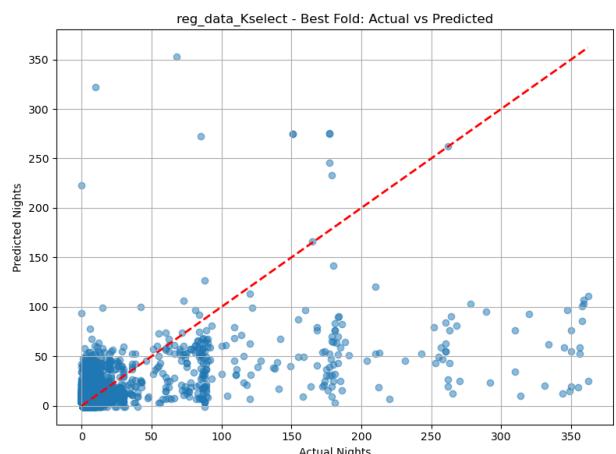


图 10: Actual Nights vs Predicted Nights

## 50 项特征提取数据:

表 13: Lasso 回归模型 50 项特征提取数据训练下五折评估指标

折数	MSE	MAE	R2
Fold 1	255.0154	6.4352	0.3708
Fold 2	4084.1204	9.2168	-11.4528
Fold 3	348.7290	5.5576	-0.0001
Fold 4	280.7685	5.8259	0.2637
Fold 5	594.3588	6.4180	-0.0024
Average	1112.5984	6.6907	-2.1642
Best	255.0154	6.4352	0.3708

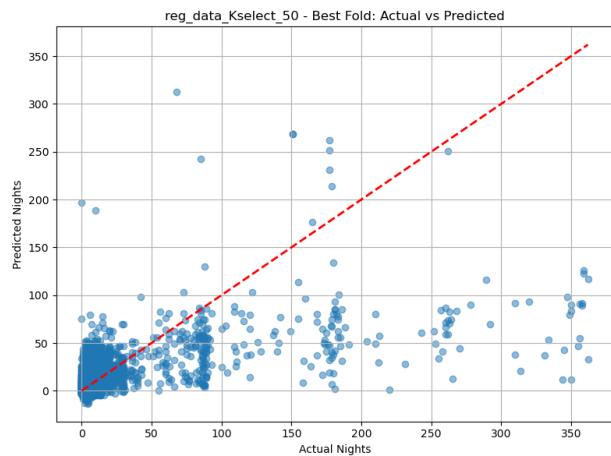


图 11: Actual Nights vs Predicted Nights

## 清洗后的 10 项特征提取数据:

表 14: Lasso 回归模型清洗后的 10 项特征提取数据训练下五折评估指标

折数	MSE	MAE	R2
Fold 1	18.3168	2.2185	0.0739
Fold 2	22.0965	2.4660	0.0073
Fold 3	18.6227	2.3356	0.0920
Fold 4	17.3055	2.1490	0.0826
Fold 5	13.5125	2.1702	0.0797
Average	17.9708	2.2679	0.0671
Best	18.6227	2.3356	0.0920

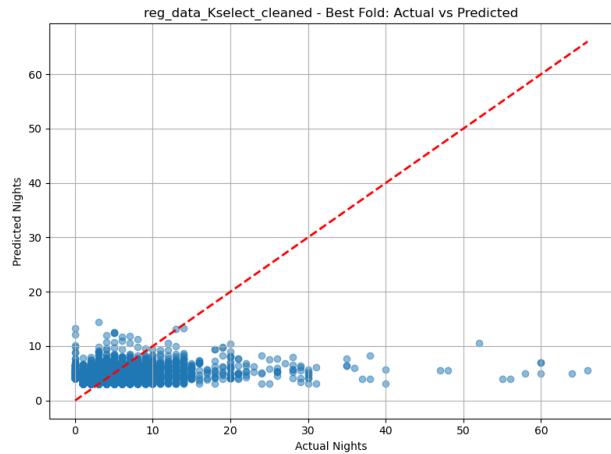


图 12: Actual Nights vs Predicted Nights

### A.3 Ridge 回归模型

原始数据：

表 15: Ridge 回归模型原始数据训练下五折评估指标

折数	MSE	MAE	R2
Fold 1	260.3425	7.2925	0.3577
Fold 2	5985.8452	11.1883	-17.2513
Fold 3	277.2545	6.4498	0.2049
Fold 4	274.2829	6.3397	0.2807
Fold 5	377.1823	6.7483	0.3639
Average	1434.9815	7.6037	-3.2088
Best	377.1823	6.7483	0.3639

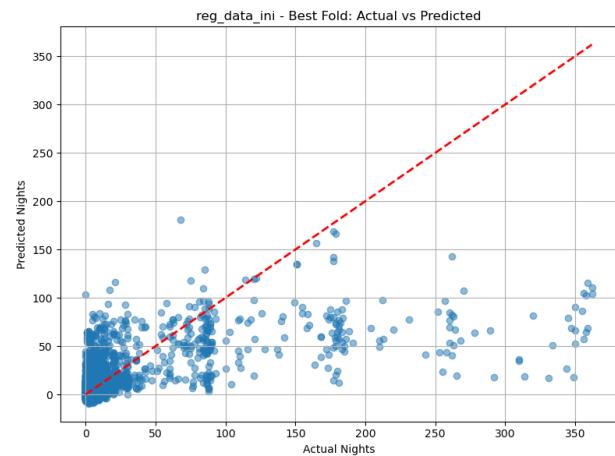


图 13: Actual Nights vs Predicted Nights

标准化数据：

表 16: Ridge 回归模型标准化数据训练下五折评估指标

折数	MSE	MAE	R2
Fold 1	262.3633	7.4057	0.3527
Fold 2	5900.5127	11.2737	-16.9911
Fold 3	277.9904	6.4628	0.2028
Fold 4	268.0924	6.3565	0.2970
Fold 5	377.1059	6.7486	0.3640
Average	1417.2129	7.6495	-3.1549
Best	377.1059	6.7486	0.3640

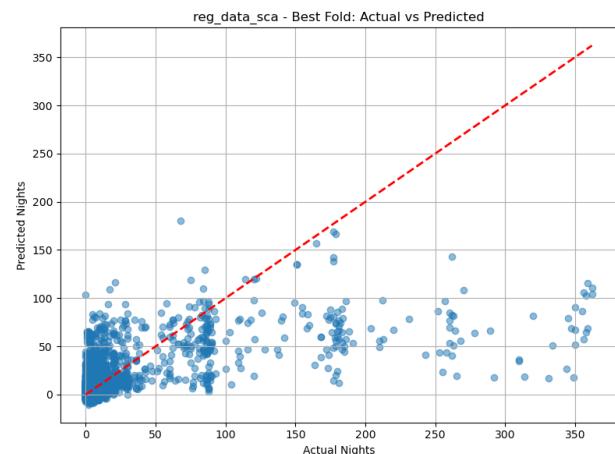


图 14: Actual Nights vs Predicted Nights

## PCA 数据:

表 17: Ridge 回归模型 PCA 数据训练下五折评估指标

折数	MSE	MAE	R2
Fold 1	312.1705	7.1230	0.2298
Fold 2	3546.0854	11.3339	-9.8123
Fold 3	307.5685	6.1770	0.1180
Fold 4	341.9265	6.1010	0.1033
Fold 5	524.3387	7.0609	0.1157
Average	1006.4179	7.5592	-1.8491
Best	312.1705	7.1230	0.2298

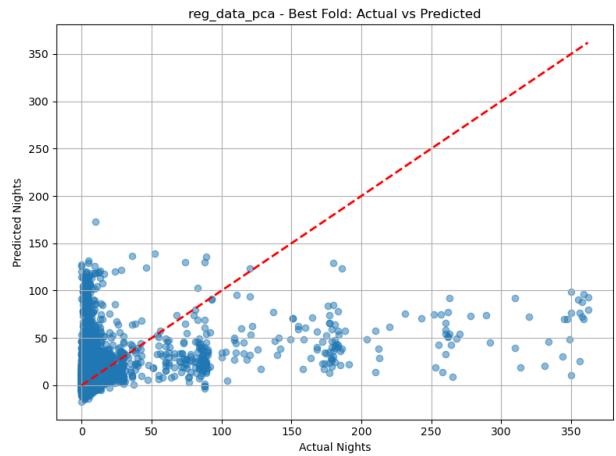


图 15: Actual Nights vs Predicted Nights

## 10 项特征提取数据:

表 18: Ridge 回归模型 10 项特征提取数据训练下五折评估指标

折数	MSE	MAE	R2
Fold 1	278.6263	6.3046	0.3125
Fold 2	2031.7844	8.2071	-5.1951
Fold 3	288.3358	5.9861	0.1731
Fold 4	272.6580	5.7056	0.2850
Fold 5	394.9882	6.3756	0.3338
Average	653.2786	6.5158	-0.8181
Best	394.9882	6.3756	0.3338

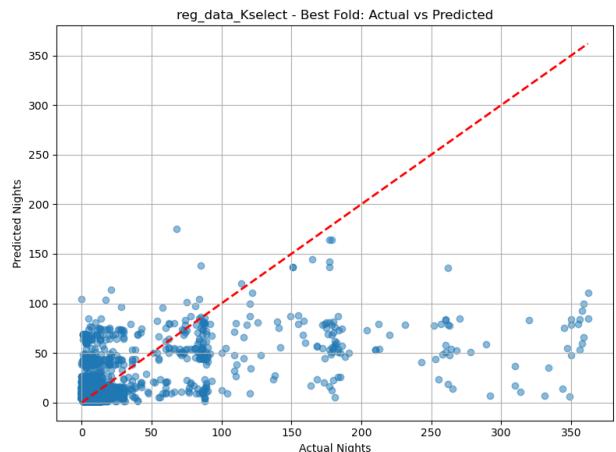


图 16: Actual Nights vs Predicted Nights

## 50 项特征提取数据:

表 19: Ridge 回归模型 50 项特征提取数据训练下五折评估指标

折数	MSE	MAE	R2
Fold 1	255.2087	6.6430	0.3703
Fold 2	4620.5263	9.6299	-13.0883
Fold 3	278.7732	5.9341	0.2005
Fold 4	269.4585	6.0046	0.2934
Fold 5	382.4843	6.5596	0.3549
Average	1161.2902	6.9542	-2.3738
Best	255.2087	6.6430	0.3703

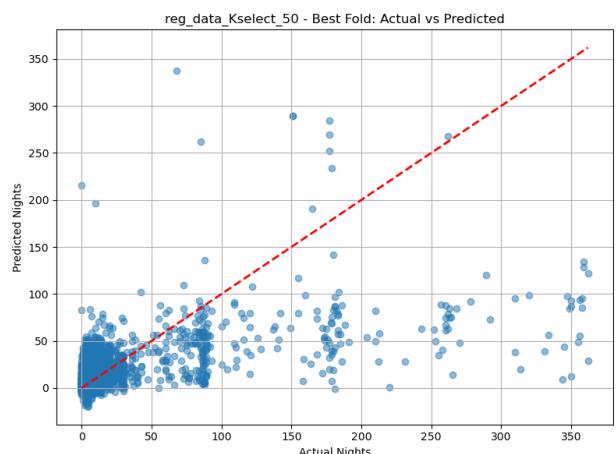


图 17: Actual Nights vs Predicted Nights

清洗后的 10 项特征提取数据：

表 20: Ridge 回归模型清洗后的 10 项特征提取数据训练下五折评估指标

折数	MSE	MAE	R2
Fold 1	18.3104	2.2219	0.0742
Fold 2	22.1839	2.4768	0.0034
Fold 3	18.6018	2.3348	0.0930
Fold 4	17.2910	2.1480	0.0834
Fold 5	13.5112	2.1709	0.0798
Average	17.9797	2.2705	0.0668
Best	18.6018	2.3348	0.0930

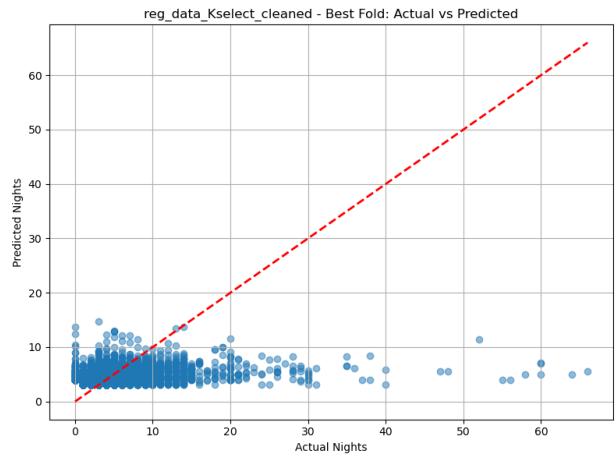


图 18: Actual Nights vs Predicted Nights

#### A.4 随机森林模型

标准化数据：

表 21: 随机森林模型标准化数据训练下五折评估指标

折数	MSE	MAE	R2
Fold 1	211.4701	4.2081	0.4782
Fold 2	304.0424	5.5388	0.0730
Fold 3	184.8663	3.8326	0.4698
Fold 4	183.9467	3.6730	0.5176
Fold 5	211.6675	3.8089	0.6430
Average	219.1986	4.2123	0.4363
Best	211.6675	3.8089	0.6430

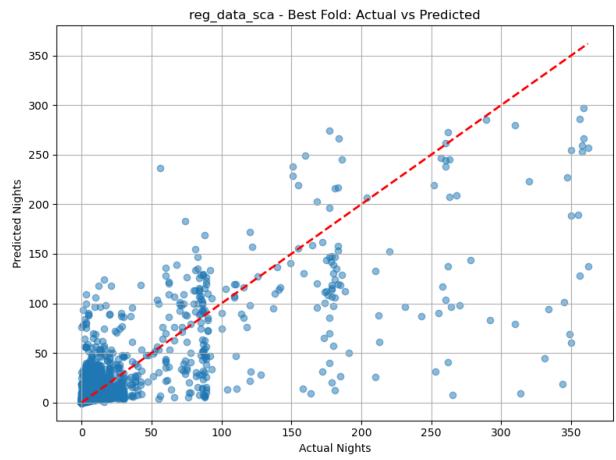


图 19: Actual Nights vs Predicted Nights

## PCA 数据:

表 22: 随机森林模型 PCA 数据训练下五折评估指标

折数	MSE	MAE	R2
Fold 1	284.2872	5.3487	0.2986
Fold 2	383.5303	6.5348	-0.1694
Fold 3	257.9379	4.7256	0.2603
Fold 4	264.6351	4.8651	0.3060
Fold 5	353.9304	5.4302	0.4031
Average	308.8642	5.3809	0.2197
Best	353.9304	5.4302	0.4031

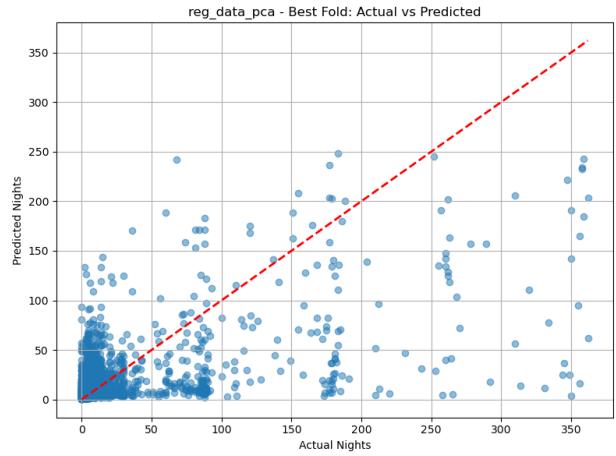


图 20: Actual Nights vs Predicted Nights

## 10 项特征提取数据:

表 23: 随机森林模型 10 项特征提取数据训练下五折评估指标

折数	MSE	MAE	R2
Fold 1	196.5421	4.1510	0.5151
Fold 2	244.8156	4.8132	0.2535
Fold 3	232.4115	4.2022	0.3335
Fold 4	203.0266	3.8641	0.4676
Fold 5	255.8492	4.3461	0.5685
Average	226.5290	4.2753	0.4276
Best	255.8492	4.3461	0.5685

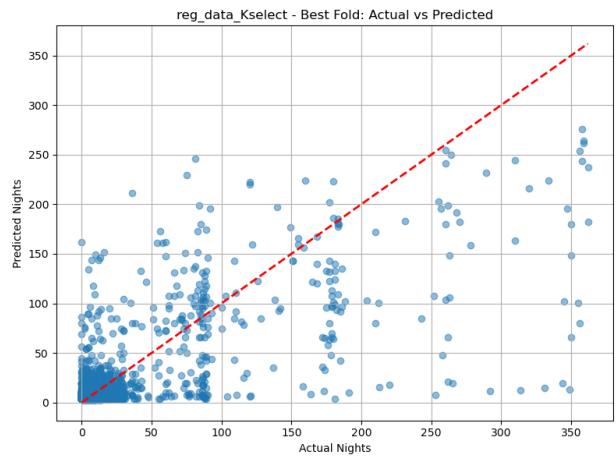


图 21: Actual Nights vs Predicted Nights

## 50 项特征提取数据:

表 24: 随机森林模型 50 项特征提取数据训练下五折评估指标

折数	MSE	MAE	R2
Fold 1	197.2213	4.2446	0.5134
Fold 2	296.4568	5.3218	0.0961
Fold 3	193.1003	3.7999	0.4462
Fold 4	183.1465	3.6573	0.5197
Fold 5	234.9864	4.1050	0.6037
Average	220.9823	4.2257	0.4358
Best	234.9864	4.1050	0.6037

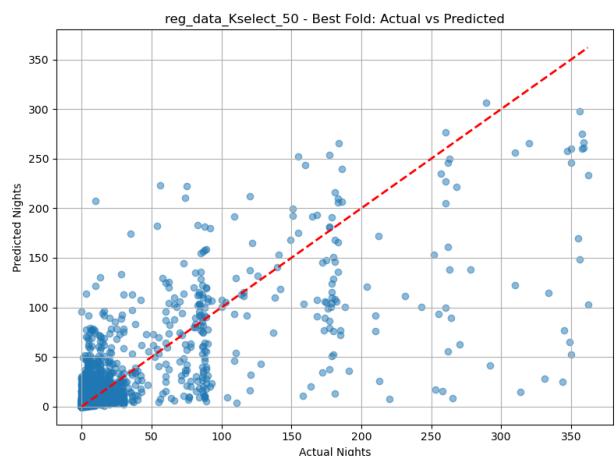


图 22: Actual Nights vs Predicted Nights

清洗后的 10 项特征提取数据：

表 25：随机森林模型清洗后的 10 项特征提取数据训练下五折评估指标

折数	MSE	MAE	R2
Fold 1	17.7608	2.1529	0.1020
Fold 2	21.3350	2.4645	0.0415
Fold 3	17.6212	2.2466	0.1408
Fold 4	16.5217	2.0615	0.1242
Fold 5	12.9729	2.0930	0.1165
Average	17.2423	2.2037	0.1050
Best	17.6212	2.2466	0.1408

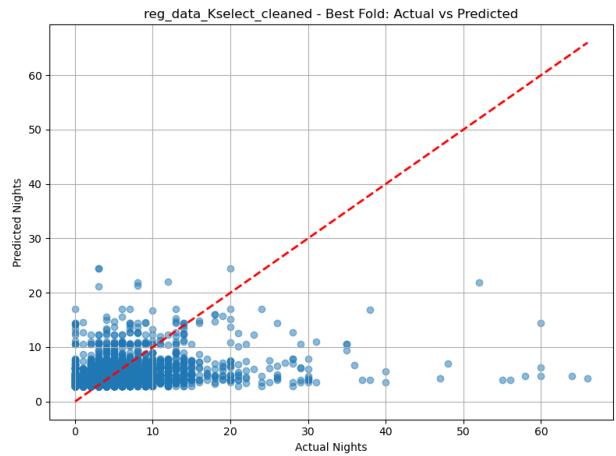


图 23: Actual Nights vs Predicted Nights

## A.5 SVR

PCA 数据：

表 26: SVR 在 PCA 数据训练下五折评估指标

折数	MSE	MAE	R2
Fold 1	364.5163	4.2061	0.1006
Fold 2	280.4139	3.8682	0.1450
Fold 3	303.5393	3.8196	0.1295
Fold 4	317.7262	3.7525	0.1668
Fold 5	476.1165	4.6908	0.1970
Average	348.4624	4.0674	0.1478
Best	476.1165	4.6908	0.1970

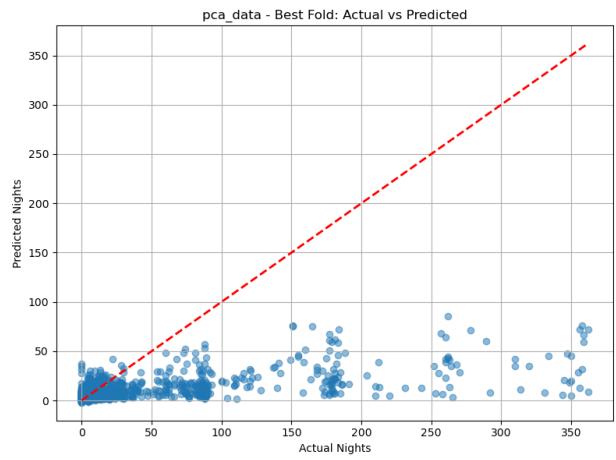


图 24: Actual Nights vs Predicted Nights

## 10 项特征提取数据:

表 27: SVR 在 10 项特征提取数据训练下五折评估指标

折数	MSE	MAE	R2
Fold 1	290.6699	4.0755	0.2828
Fold 2	258.7043	4.1154	0.2112
Fold 3	274.7173	3.9747	0.2122
Fold 4	290.2829	3.7802	0.2388
Fold 5	379.6527	4.4638	0.3597
Average	298.8054	4.0819	0.2609
Best	379.6527	4.4638	0.3597

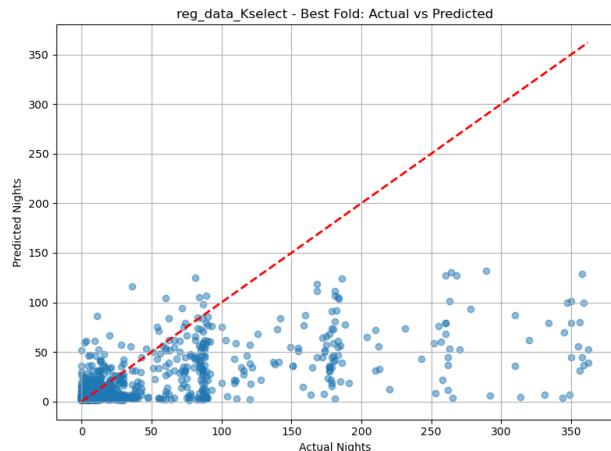


图 25: Actual Nights vs Predicted Nights

## 50 项特征提取数据:

表 28: SVR 在 50 项特征提取数据训练下五折评估指标

折数	MSE	MAE	R2
Fold 1	355.2848	4.0209	0.1234
Fold 2	275.6529	3.9686	0.1595
Fold 3	300.5682	3.8578	0.1380
Fold 4	327.3428	3.7428	0.1416
Fold 5	468.6323	4.5820	0.2096
Average	345.4962	4.0344	0.1544
Best	468.6323	4.5820	0.2096

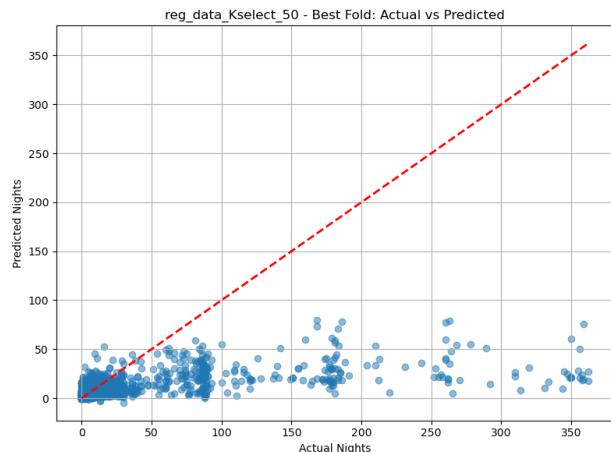


图 26: Actual Nights vs Predicted Nights

## 清洗后的 10 项特征提取数据:

表 29: SVR 清洗后的 10 项特征提取数据训练下五折评估指标

折数	MSE	MAE	R2
Fold 1	18.7741	2.1039	0.0508
Fold 2	22.6115	2.3760	-0.0158
Fold 3	19.0037	2.1750	0.0734
Fold 4	17.6689	2.0315	0.0633
Fold 5	13.7260	2.0511	0.0652
Average	18.3568	2.1475	0.0474
Best	19.0037	2.1750	0.0734

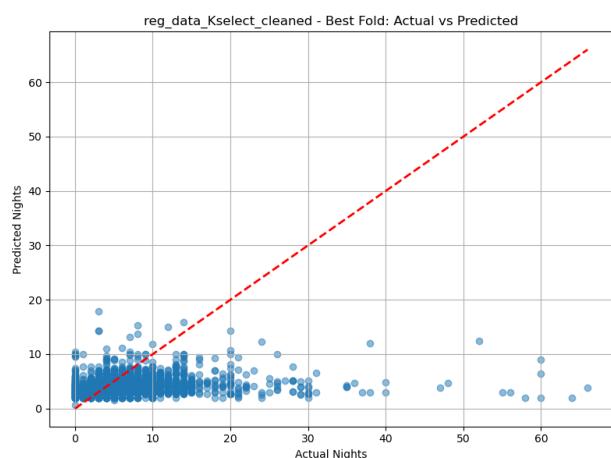


图 27: Actual Nights vs Predicted Nights

## A2 分类任务结果

由于数据图表实在太多，故仅仅在此展现平均之后的数据，就不再展示每一折的评估指标的结果。

### A2.1 KNN

原始数据

表 30: KNN 在原始数据训练下各项评估指标均值

Accuracy	Precision_macro	Recall_macro	F1_macro
0.6989	0.4520	0.3482	0.3647
AUC	Precision_weighted	Recall_weighted	F1_weighted
0.6707	0.6527	0.6989	0.6646

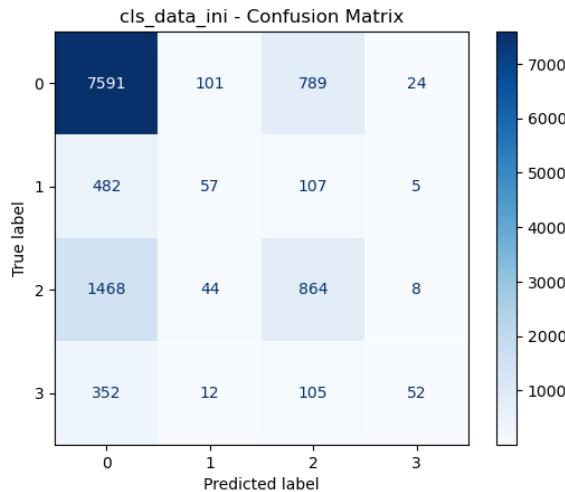


图 28: 混淆矩阵

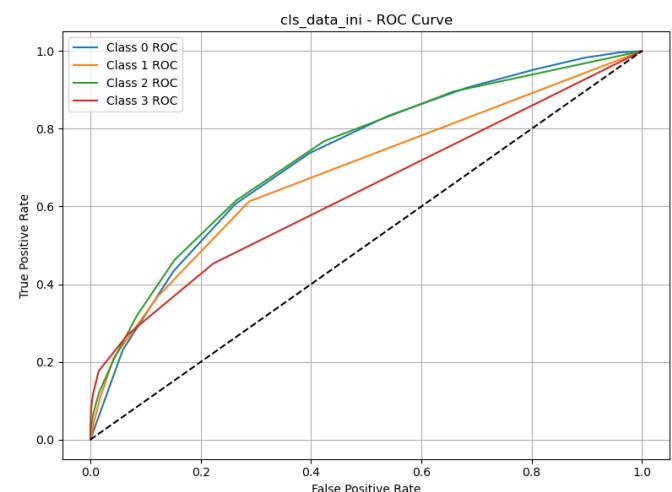


图 29: ROC 曲线

标准化数据

表 31: KNN 在标准化数据训练下各项评估指标均值

Accuracy	Precision_macro	Recall_macro	F1_macro
0.8290	0.6873	0.5133	0.5292
AUC	Precision_weighted	Recall_weighted	F1_weighted
0.8297	0.8149	0.8290	0.8086

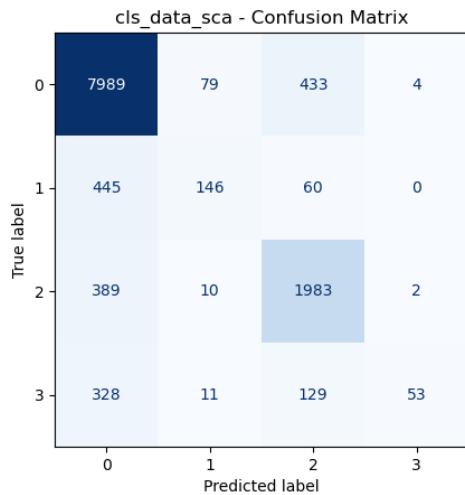


图 30: 混淆矩阵

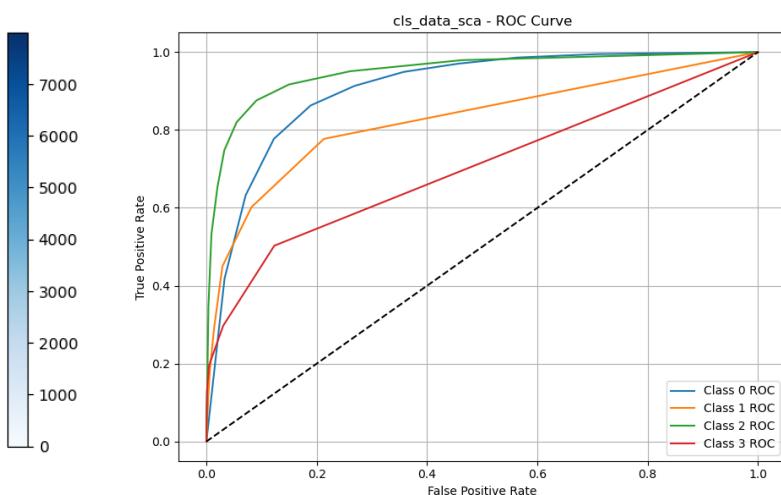


图 31: ROC 曲线

## LDA 降维数据

表 32: KNN 在 LDA 降维数据训练下各项评估指标均值

Accuracy	Precision_macro	Recall_macro	F1_macro
0.8748	0.7251	0.6478	0.6750
AUC	Precision_weighted	Recall_weighted	F1_weighted
0.8954	0.8659	0.8748	0.8683

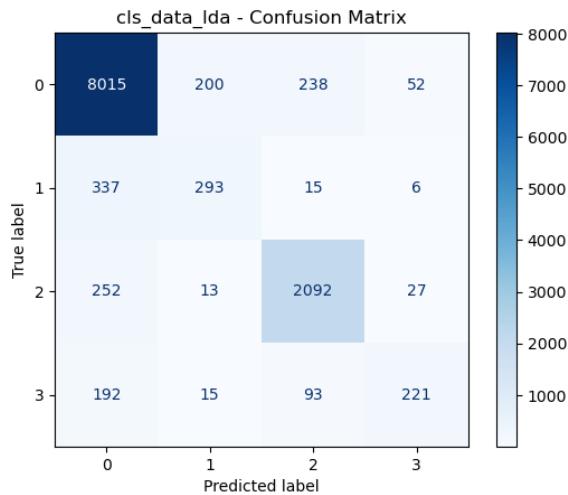


图 32: 混淆矩阵

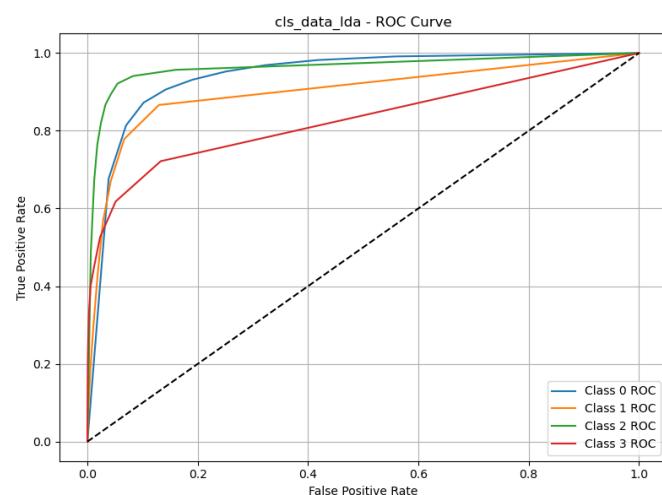


图 33: ROC 曲线

## 10 项特征提取数据

表 33: KNN 在 10 项特征提取数据训练下各项评估指标均值

Accuracy	Precision_macro	Recall_macro	F1_macro
0.8116	0.6130	0.6181	0.6073
AUC	Precision_weighted	Recall_weighted	F1_weighted
0.8466	0.8196	0.8116	0.8137

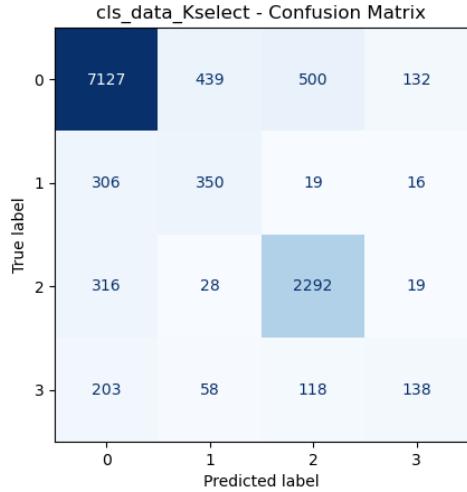


图 34: 混淆矩阵

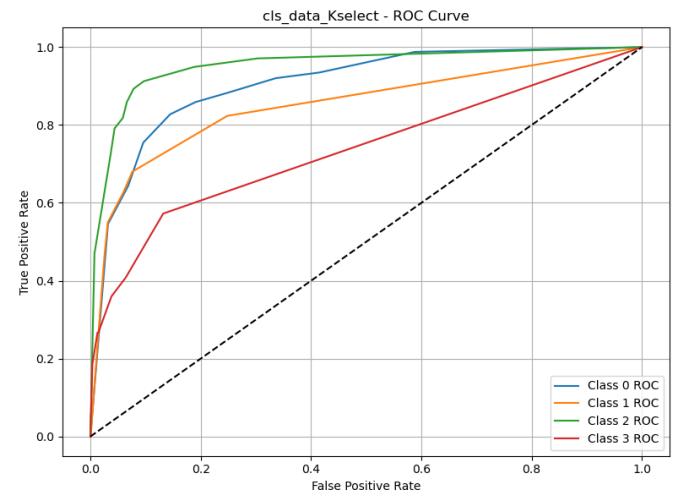


图 35: ROC 曲线

## 50 项特征提取数据

表 34: KNN 在 50 项特征提取数据训练下各项评估指标均值

Accuracy	Precision_macro	Recall_macro	F1_macro
0.8740	0.7401	0.6211	0.6580
AUC	Precision_weighted	Recall_weighted	F1_weighted
0.8779	0.8624	0.8740	0.8640

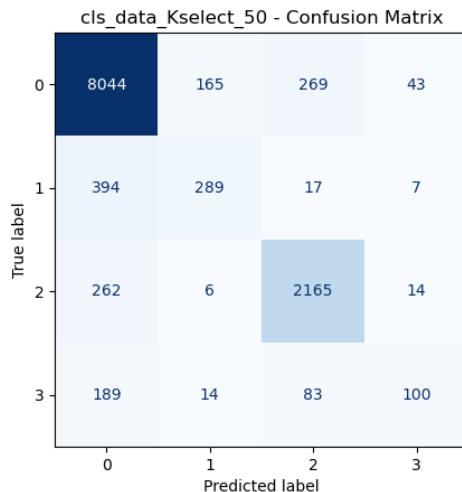


图 36: 混淆矩阵

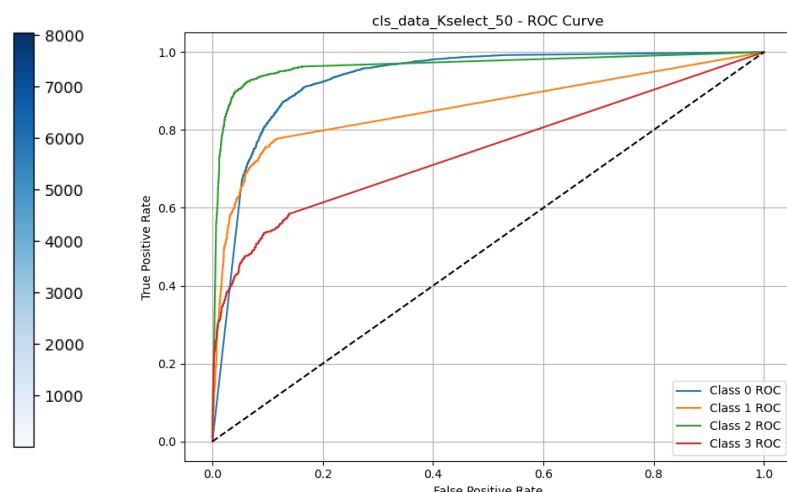


图 37: ROC 曲线

### 清洗后的 10 项特征提取数据

表 35: KNN 在清洗后的 10 项特征提取数据训练下各项评估指标均值

Accuracy	Precision_macro	Recall_macro	F1_macro
0.8482	0.4754	0.4938	0.4721
AUC	Precision_weighted	Recall_weighted	F1_weighted
0.8005	0.8500	0.8482	0.8478

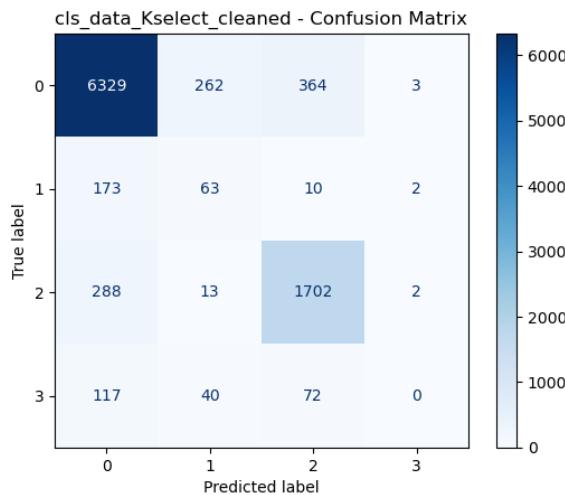


图 38: 混淆矩阵

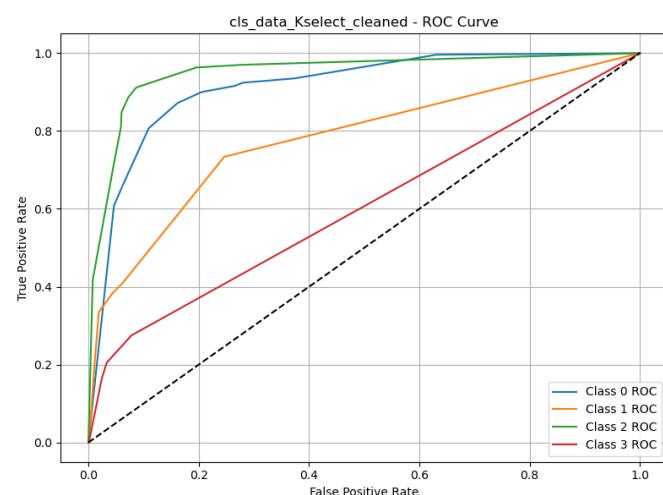


图 39: ROC 曲线

## A2.2 Logistic

原始数据

表 36: Logistic 在原始数据训练下各项评估指标均值

Accuracy	Precision_macro	Recall_macro	F1_macro
0.8803	0.7750	0.6353	0.6764
AUC	Precision_weighted	Recall_weighted	F1_weighted
0.9332	0.8716	0.8803	0.8706

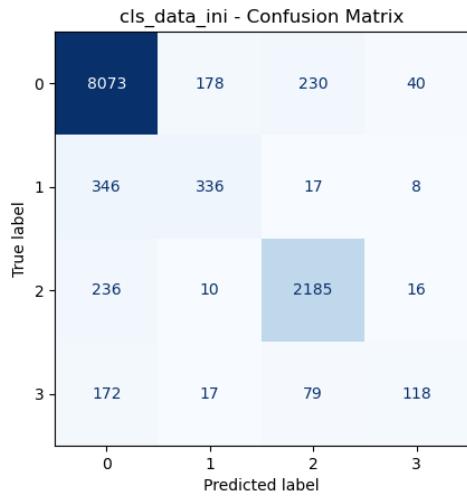


图 40: 混淆矩阵

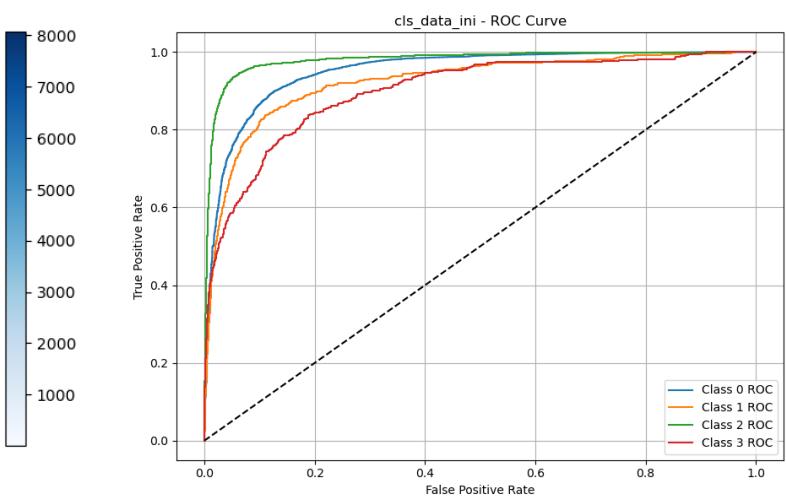


图 41: ROC 曲线

标准化数据

表 37: Logistic 在标准化数据训练下各项评估指标均值

Accuracy	Precision_macro	Recall_macro	F1_macro
0.8790	0.7716	0.6322	0.6725
AUC	Precision_weighted	Recall_weighted	F1_weighted
0.9323	0.8704	0.8790	0.8691

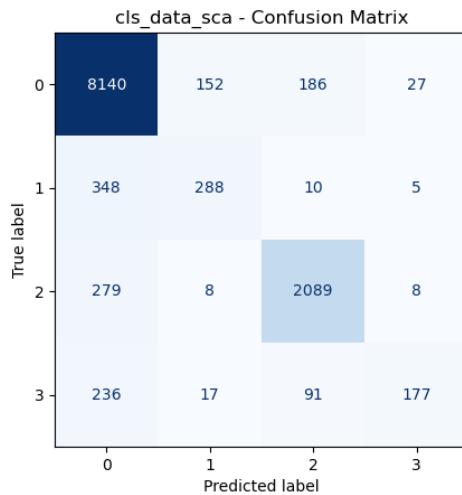


图 42: 混淆矩阵

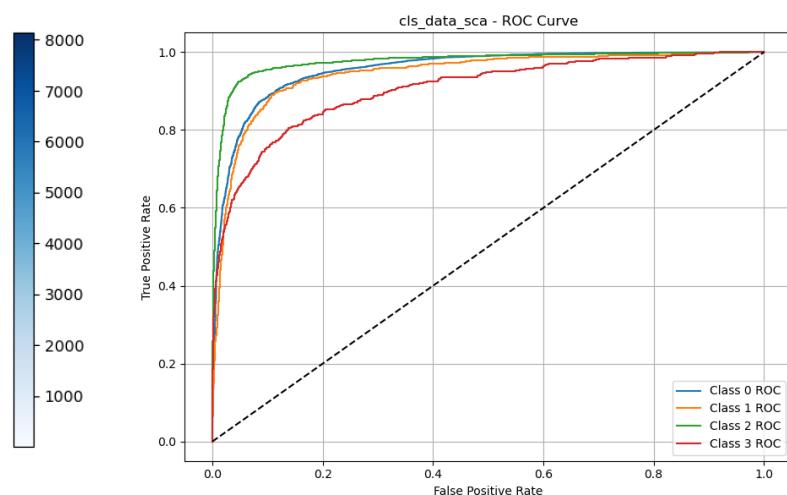


图 43: ROC 曲线

## LDA 降维数据

表 38: Logistic 在 LDA 降维数据训练下各项评估指标均值

Accuracy	Precision_macro	Recall_macro	F1_macro
0.8809	0.7675	0.6418	0.6801
AUC	Precision_weighted	Recall_weighted	F1_weighted
0.9426	0.8721	0.8809	0.8721

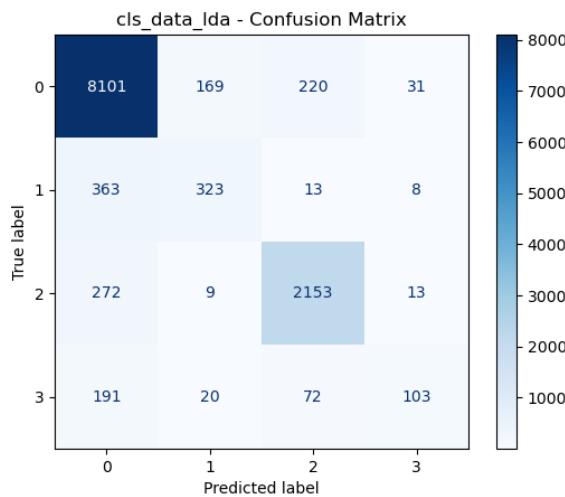


图 44: 混淆矩阵

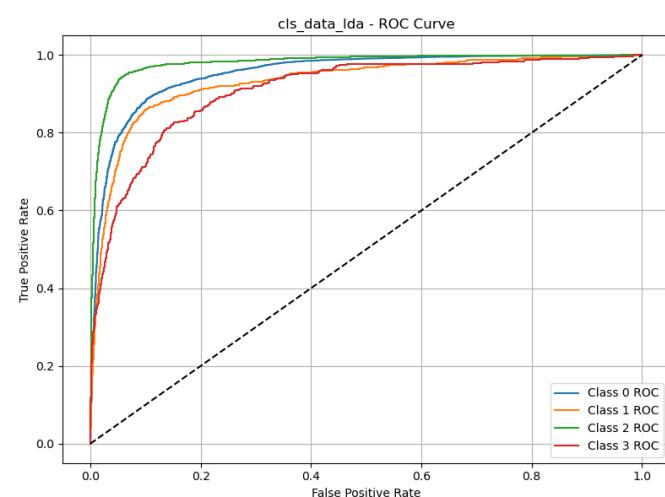


图 45: ROC 曲线

## 10 项特征提取数据

表 39: Logistic 在 10 项特征提取数据训练下各项评估指标均值

Accuracy	Precision_macro	Recall_macro	F1_macro
0.8486	0.7309	0.5551	0.6013
AUC	Precision_weighted	Recall_weighted	F1_weighted
0.9018	0.8370	0.8486	0.8327

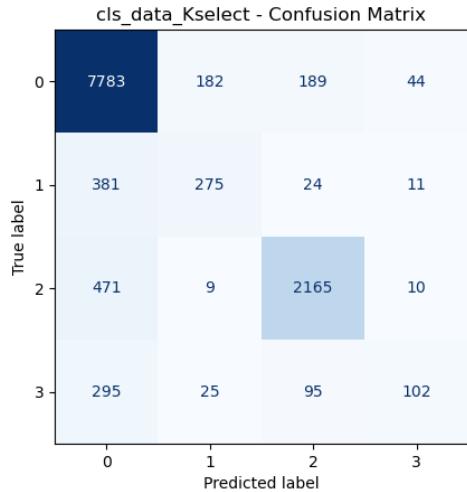


图 46: 混淆矩阵

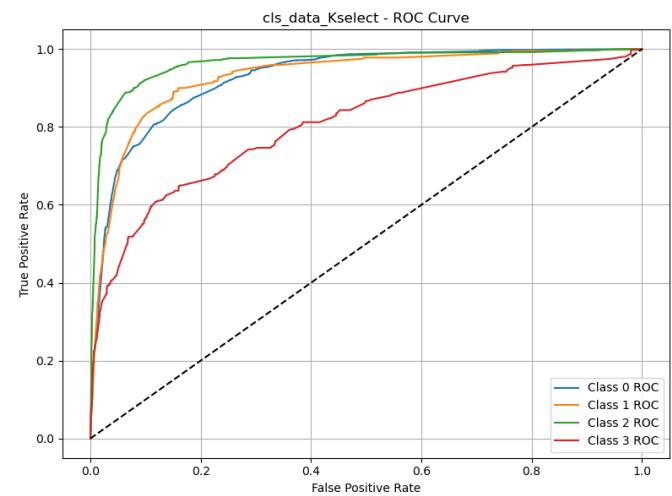


图 47: ROC 曲线

## 50 项特征提取数据

表 40: Logistic 在 50 项特征提取数据训练下各项评估指标均值

Accuracy	Precision_macro	Recall_macro	F1_macro
0.8812	0.7658	0.6315	0.6700
AUC	Precision_weighted	Recall_weighted	F1_weighted
0.9269	0.8715	0.8812	0.8713

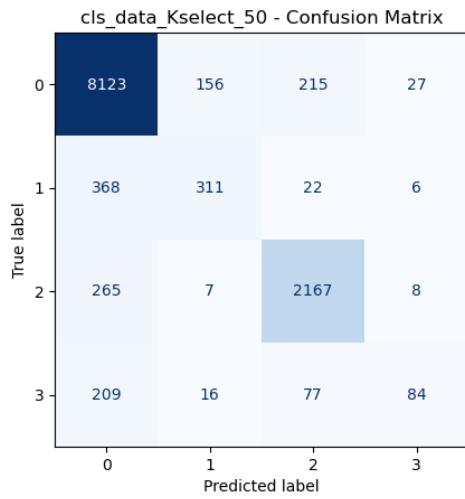


图 48: 混淆矩阵

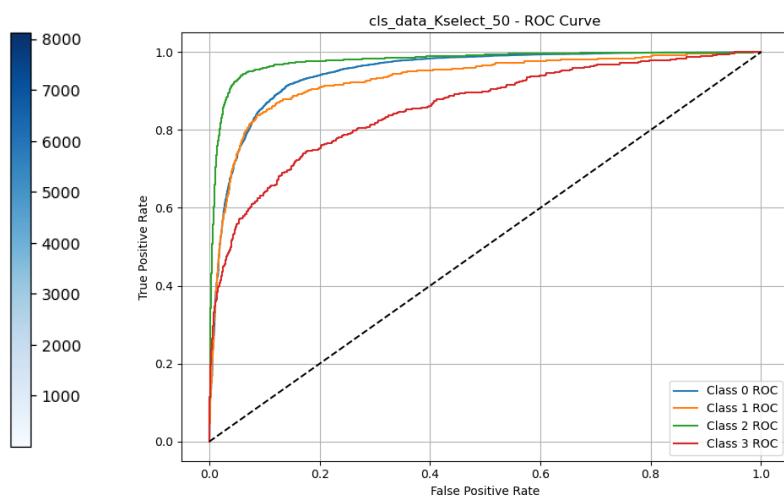


图 49: ROC 曲线

### 清洗后的 10 项特征提取数据

表 41: Logistic 在清洗后的 10 项特征提取数据训练下各项评估指标均值

	Accuracy	Precision_macro	Recall_macro	F1_macro
	0.8820	0.5310	0.4272	0.4392
AUC		Precision_weighted	Recall_weighted	F1_weighted
	0.8628	0.8511	0.8820	0.8587

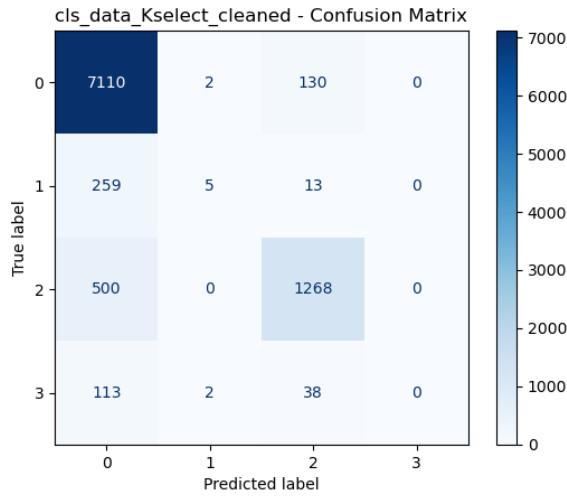


图 50: 混淆矩阵

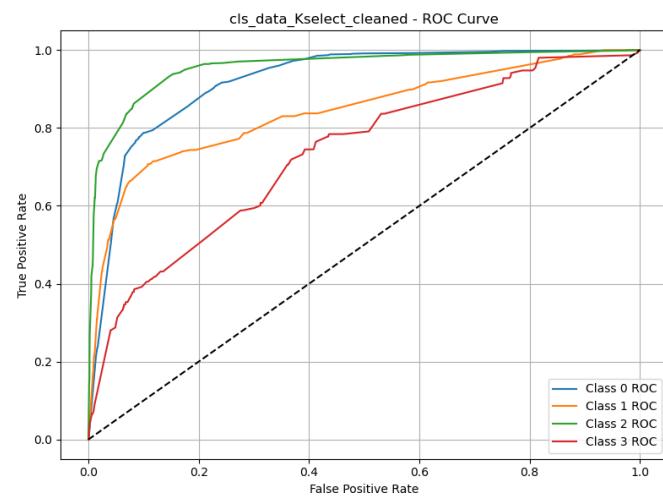


图 51: ROC 曲线

### A2.3 LDA

原始数据

表 42: LDA 在原始数据训练下各项评估指标均值

Accuracy	Precision_macro	Recall_macro	F1_macro
0.8596	0.6873	0.7091	0.6948
AUC	Precision_weighted	Recall_weighted	F1_weighted
0.9244	0.8674	0.8596	0.8624

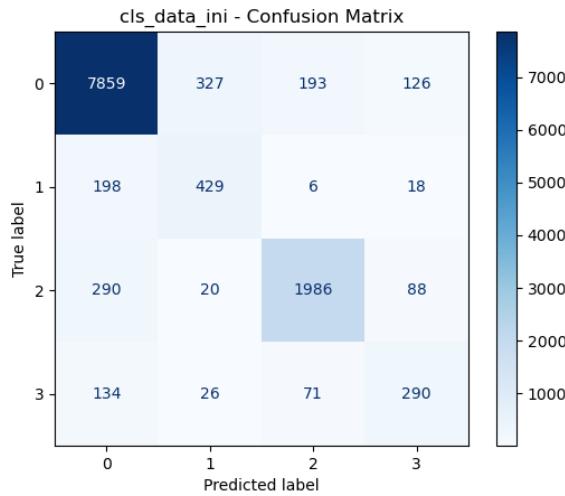


图 52: 混淆矩阵

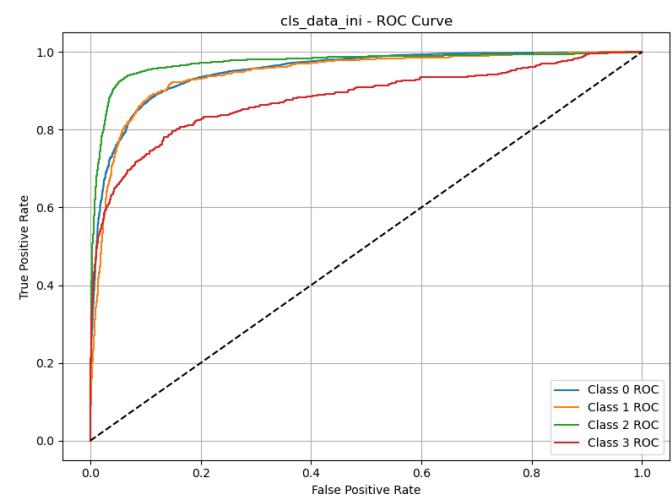


图 53: ROC 曲线

标准化数据

表 43: LDA 在标准化数据训练下各项评估指标均值

Accuracy	Precision_macro	Recall_macro	F1_macro
0.8680	0.6958	0.7126	0.7018
AUC	Precision_weighted	Recall_weighted	F1_weighted
0.9300	0.8735	0.8680	0.8700

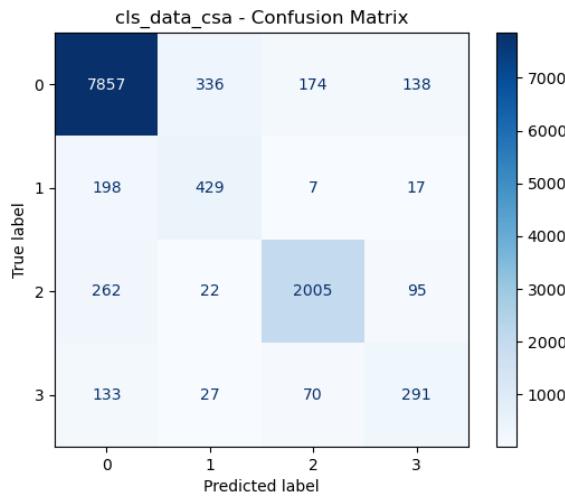


图 54: 混淆矩阵

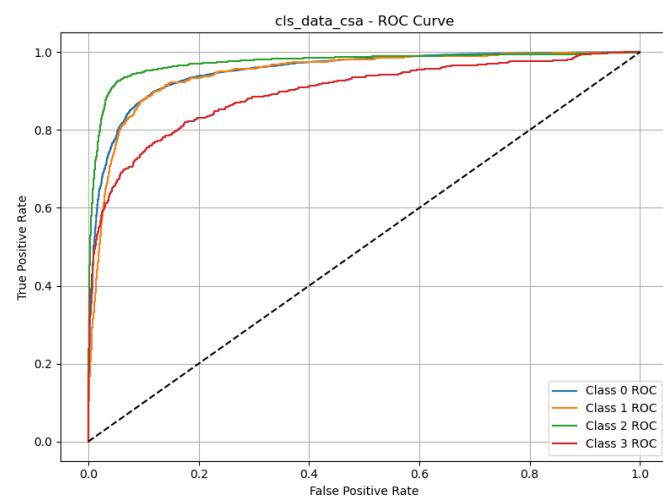


图 55: ROC 曲线

## LDA 降维数据

表 44: LDA 在 LDA 降维数据训练下各项评估指标均值

Accuracy	Precision_macro	Recall_macro	F1_macro
0.8706	0.7039	0.7178	0.7091
AUC	Precision_weighted	Recall_weighted	F1_weighted
0.9327	0.8745	0.8706	0.8721

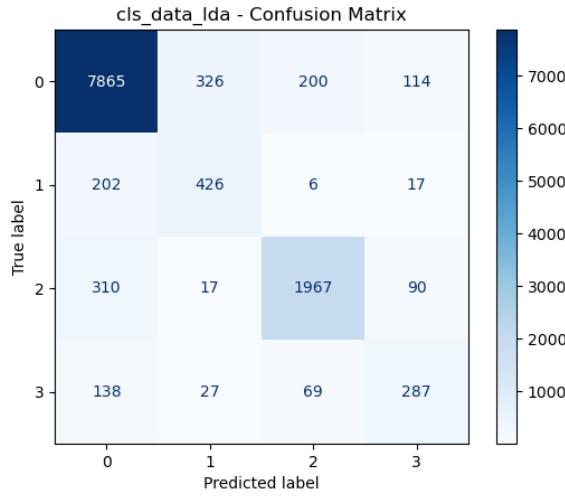


图 56: 混淆矩阵

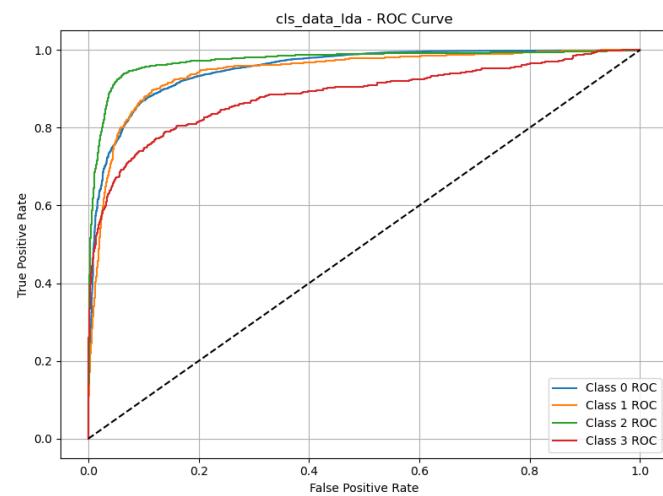


图 57: ROC 曲线

## 10 项特征提取数据

表 45: LDA 在 10 项特征提取数据训练下各项评估指标均值

Accuracy	Precision_macro	Recall_macro	F1_macro
0.8209	0.6414	0.6516	0.6391
AUC	Precision_weighted	Recall_weighted	F1_weighted
0.8896	0.8345	0.8209	0.8240

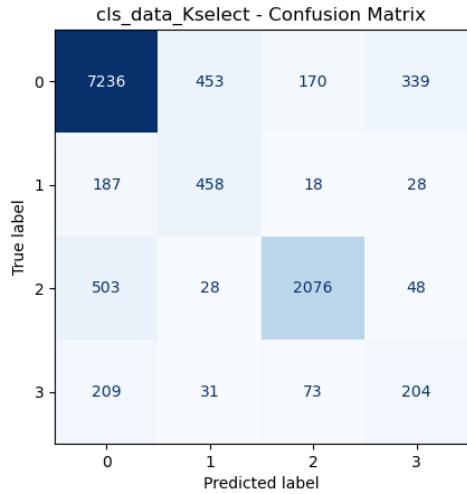


图 58: 混淆矩阵

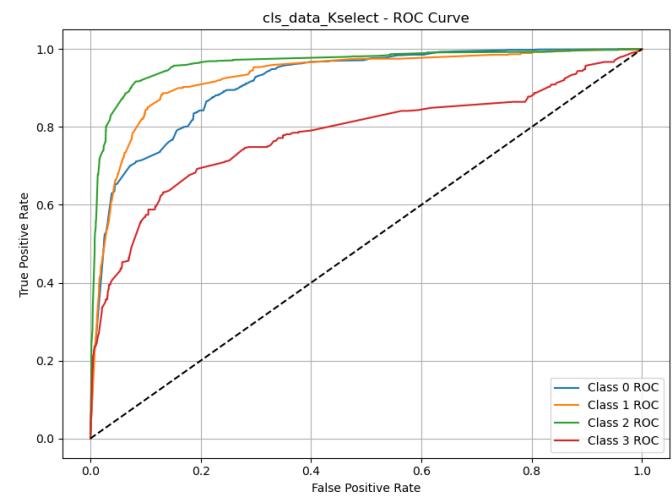


图 59: ROC 曲线

## 50 项特征提取数据

表 46: LDA 在 50 项特征提取数据训练下各项评估指标均值

Accuracy	Precision_macro	Recall_macro	F1_macro
0.8565	0.6745	0.7065	0.6866
AUC	Precision_weighted	Recall_weighted	F1_weighted
0.9182	0.8666	0.8565	0.8604

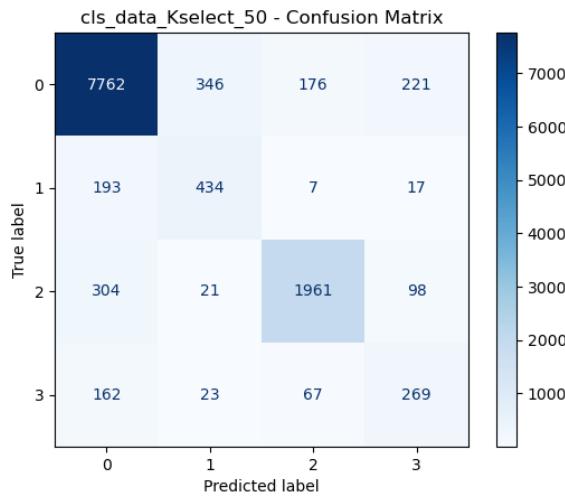


图 60: 混淆矩阵

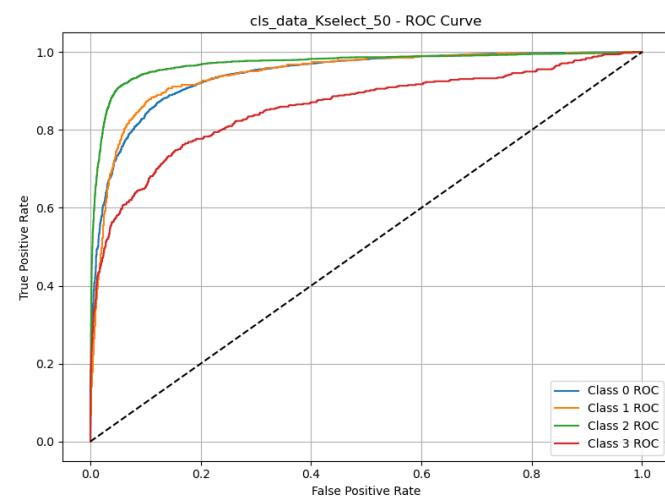


图 61: ROC 曲线

### 清洗后的 10 项特征提取数据

表 47: LDA 在清洗后的 10 项特征提取数据训练下各项评估指标均值

Accuracy	Precision_macro	Recall_macro	F1_macro
0.8381	0.5515	0.5573	0.5061
AUC	Precision_weighted	Recall_weighted	F1_weighted
0.8465	0.8615	0.8381	0.8429

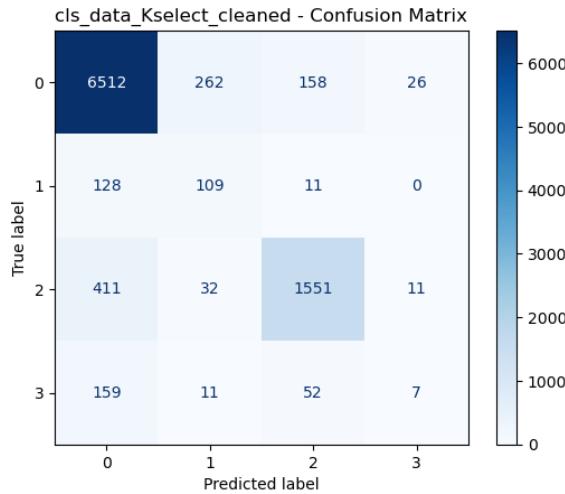


图 62: 混淆矩阵

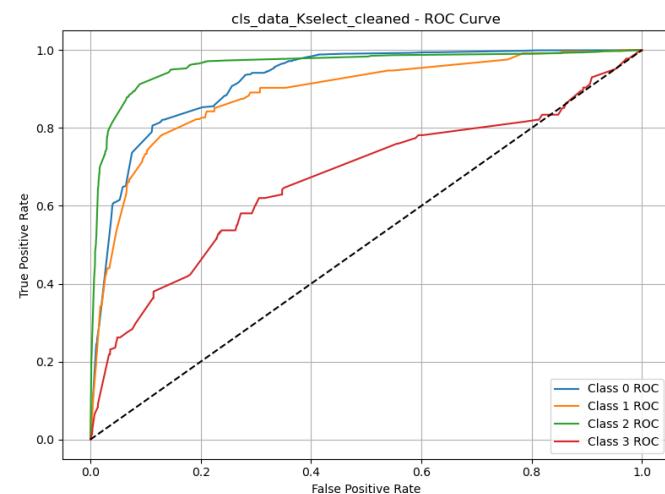


图 63: ROC 曲线

## A2.4 贝叶斯

原始数据

表 48: 贝叶斯在原始数据训练下各项评估指标均值

Accuracy	Precision_macro	Recall_macro	F1_macro
0.4029	0.3816	0.2861	0.2047
AUC	Precision_weighted	Recall_weighted	F1_weighted
0.6225	0.5619	0.4029	0.3191

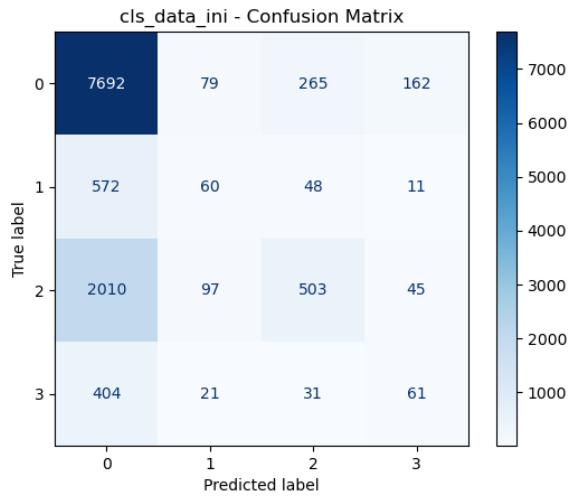


图 64: 混淆矩阵

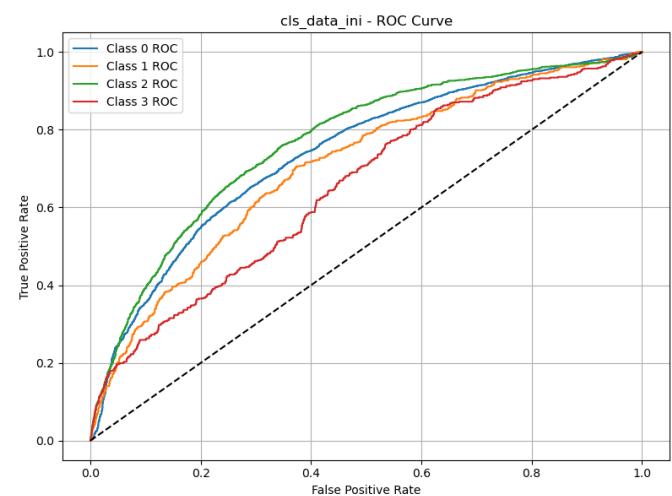


图 65: ROC 曲线

标准化数据

表 49: 贝叶斯在标准化数据训练下各项评估指标均值

Accuracy	Precision_macro	Recall_macro	F1_macro
0.0805	0.3416	0.3662	0.0879
AUC	Precision_weighted	Recall_weighted	F1_weighted
0.7381	0.5981	0.0805	0.0357

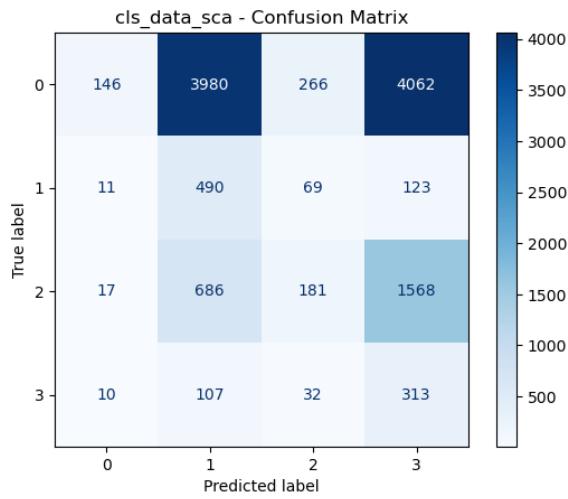


图 66: 混淆矩阵

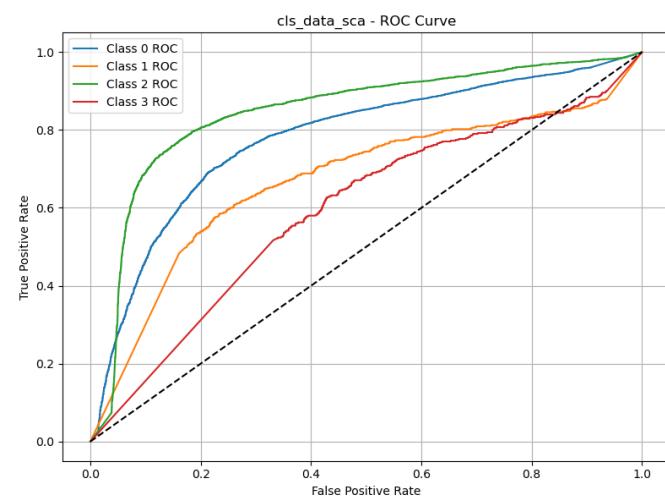


图 67: ROC 曲线

表 50: 贝叶斯在 LDA 降维数据训练下各项评估指标均值

	Accuracy	Precision_macro	Recall_macro	F1_macro
	0.8719	0.6982	0.7134	0.7048
AUC		Precision_weighted	Recall_weighted	F1_weighted
	0.9309	0.8753	0.8719	0.8734

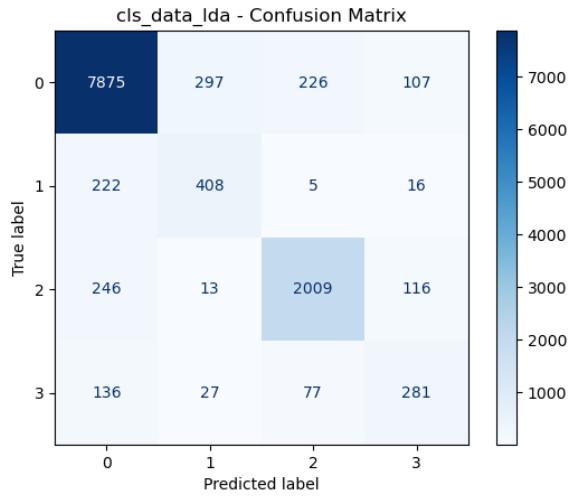


图 68: 混淆矩阵

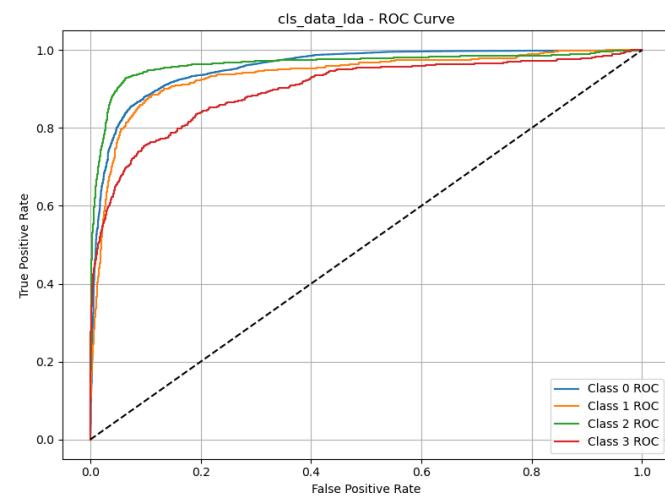


图 69: ROC 曲线

## 10 项特征提取数据

表 51: 贝叶斯在 10 项特征提取数据训练下各项评估指标均值

	Accuracy	Precision_macro	Recall_macro	F1_macro
	0.7959	0.5889	0.6948	0.6242
AUC		Precision_weighted	Recall_weighted	F1_weighted
	0.8896	0.8357	0.7959	0.8092

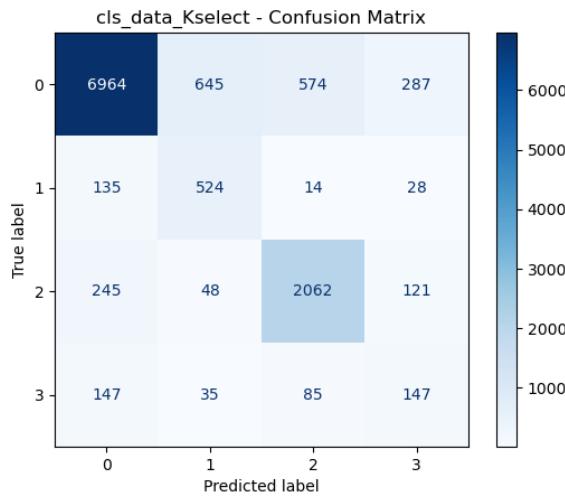


图 70: 混淆矩阵

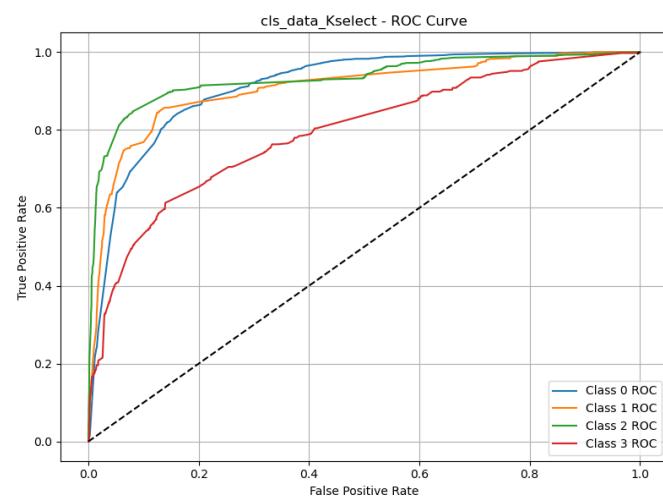


图 71: ROC 曲线

## 50 项特征提取数据

表 52: 贝叶斯在 50 项特征提取数据训练下各项评估指标均值

Accuracy	Precision_macro	Recall_macro	F1_macro
0.7959	0.5936	0.7258	0.6249
AUC	Precision_weighted	Recall_weighted	F1_weighted
0.9066	0.8703	0.7959	0.8216

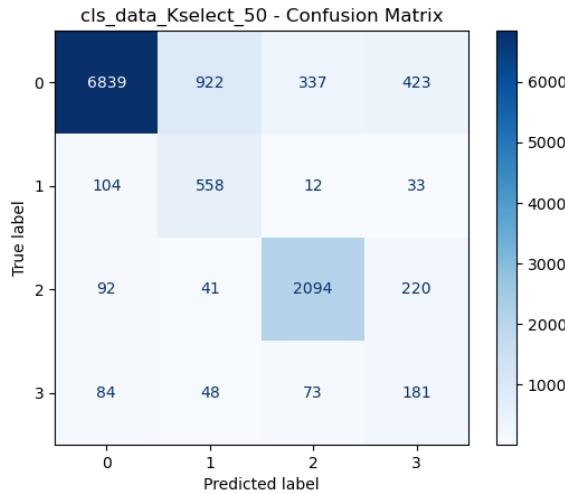


图 72: 混淆矩阵

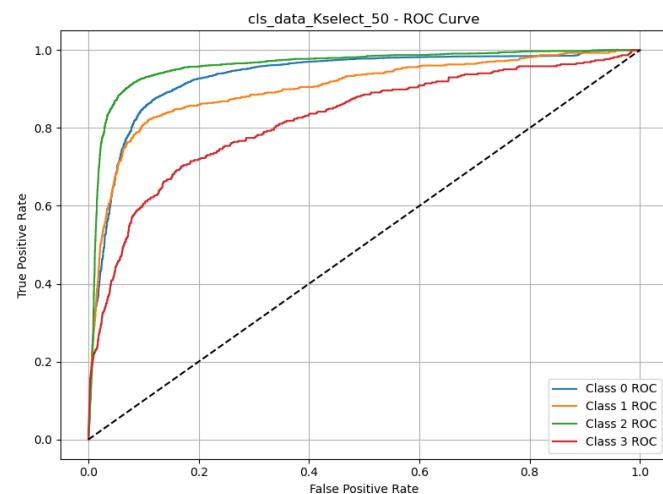


图 73: ROC 曲线

## 清洗后的 10 项特征提取数据

表 53: 贝叶斯在清洗后的 10 项特征提取数据训练下各项评估指标均值

Accuracy	Precision_macro	Recall_macro	F1_macro
0.8262	0.4799	0.5649	0.4985
AUC	Precision_weighted	Recall_weighted	F1_weighted
0.8580	0.8582	0.8262	0.8380

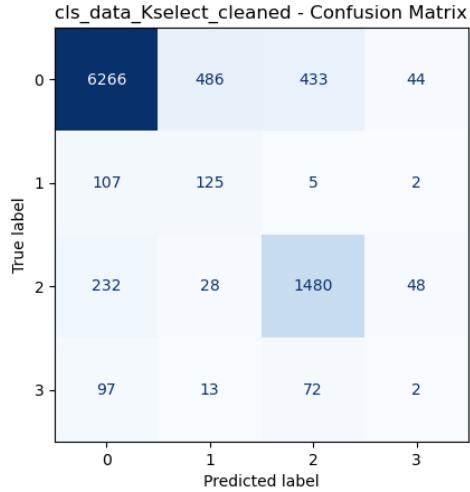


图 74: 混淆矩阵

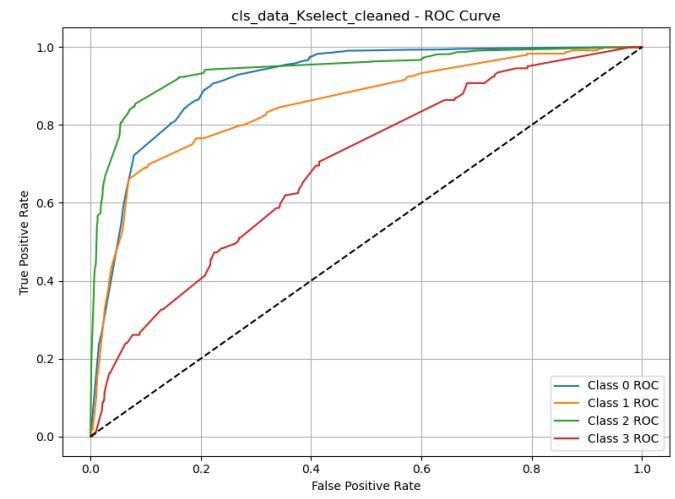


图 75: ROC 曲线

## A2.5 随机森林

表 54: 随机森林在 LDA 降维数据训练下各项评估指标均值

Accuracy	Precision_macro	Recall_macro	F1_macro
0.8838	0.7588	0.6535	0.6881
AUC	Precision_weighted	Recall_weighted	F1_weighted
0.9417	0.8746	0.8838	0.8760

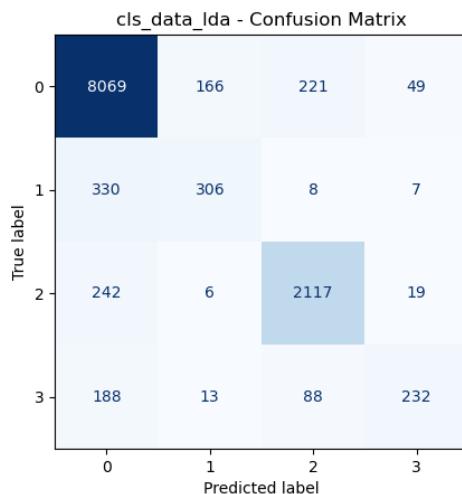


图 76: 混淆矩阵

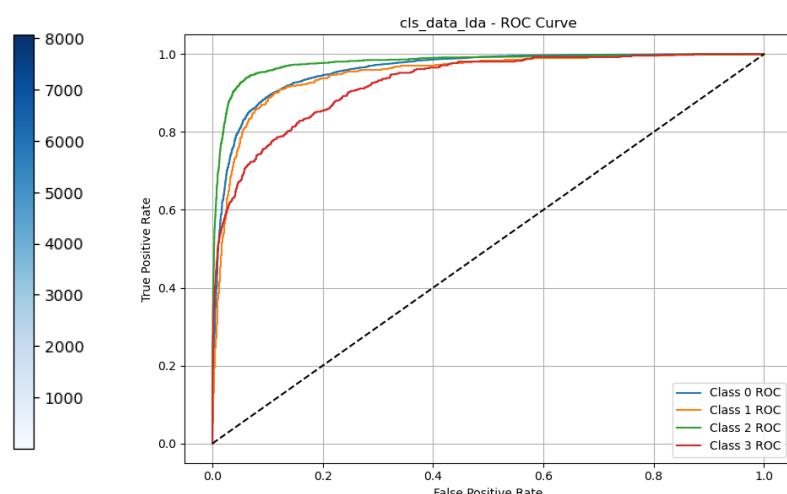


图 77: ROC 曲线

## 10 项特征提取数据

表 55: 随机森林在 10 项特征提取数据训练下各项评估指标均值

Accuracy	Precision_macro	Recall_macro	F1_macro
0.8501	0.7217	0.5814	0.6264
AUC	Precision_weighted	Recall_weighted	F1_weighted
0.9008	0.8390	0.8501	0.8379

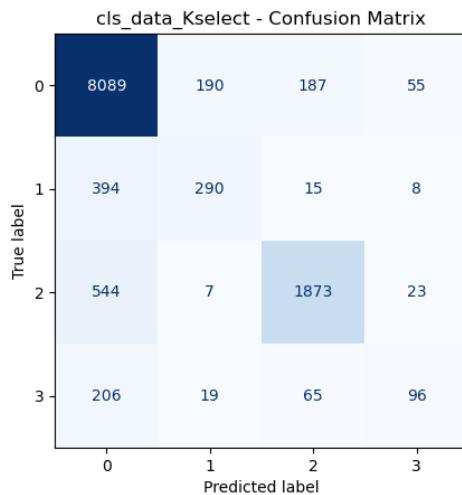


图 78: 混淆矩阵

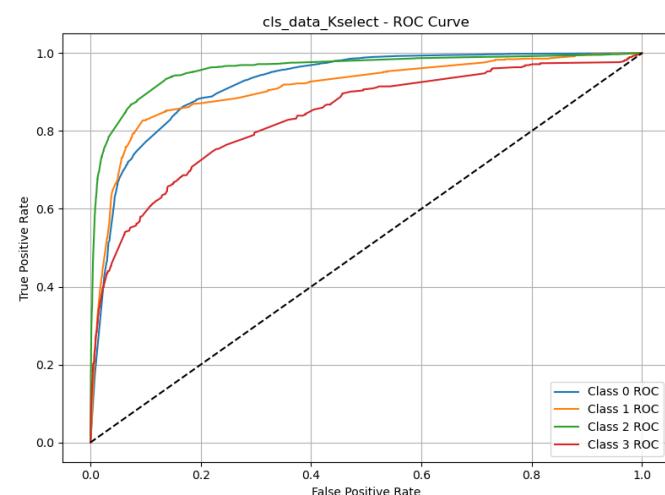


图 79: ROC 曲线

## 50 项特征提取数据

表 56: 随机森林在 50 项特征提取数据训练下各项评估指标均值

Accuracy	Precision_macro	Recall_macro	F1_macro
0.8830	0.8044	0.6210	0.6670
AUC	Precision_weighted	Recall_weighted	F1_weighted
0.9347	0.8752	0.8830	0.8709

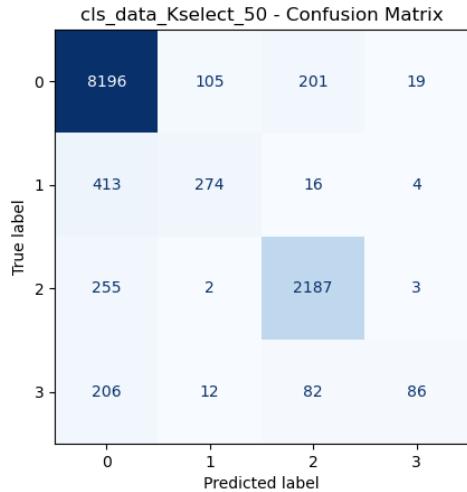


图 80: 混淆矩阵

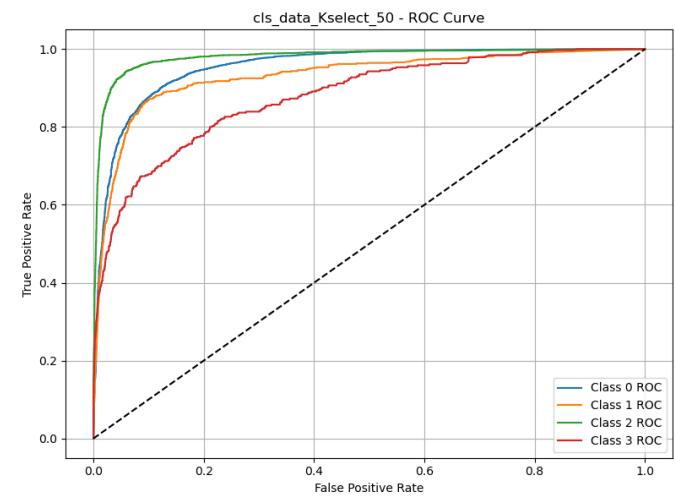


图 81: ROC 曲线

## 清洗后的 10 项特征提取数据

表 57: 随机森林在清洗后的 10 项特征提取数据训练下各项评估指标均值

Accuracy	Precision_macro	Recall_macro	F1_macro
0.8828	0.4643	0.4218	0.4298
AUC	Precision_weighted	Recall_weighted	F1_weighted
0.8692	0.8440	0.8828	0.8582

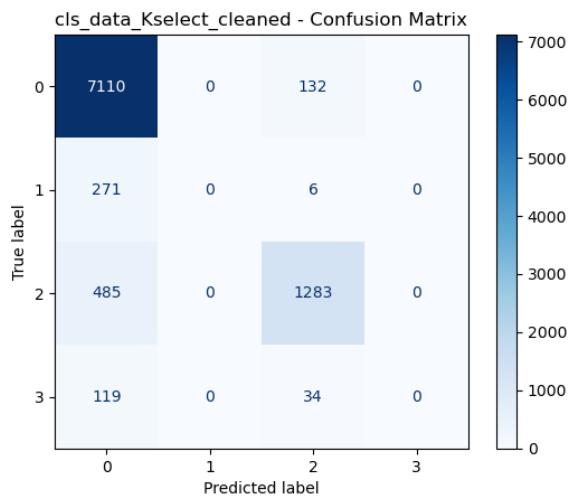


图 82: 混淆矩阵

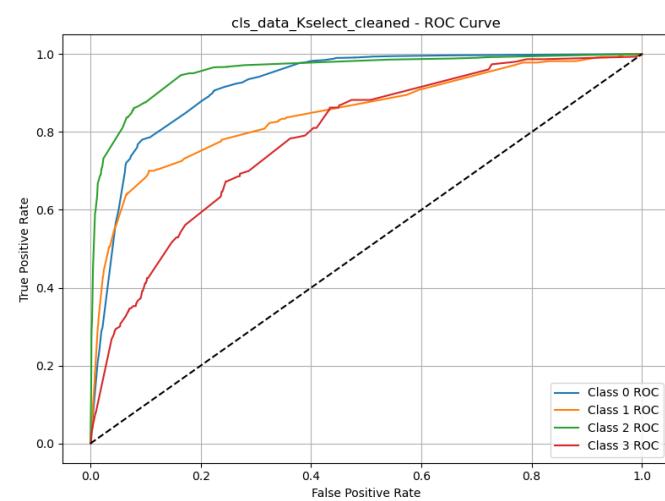


图 83: ROC 曲线

## A2.6 SVM

表 58: SVM 在 LDA 降维数据训练下各项评估指标均值

Accuracy	Precision_macro	Recall_macro	F1_macro
0.8852	0.7722	0.6497	0.6864
AUC	Precision_weighted	Recall_weighted	F1_weighted
0.8882	0.8765	0.8852	0.8766

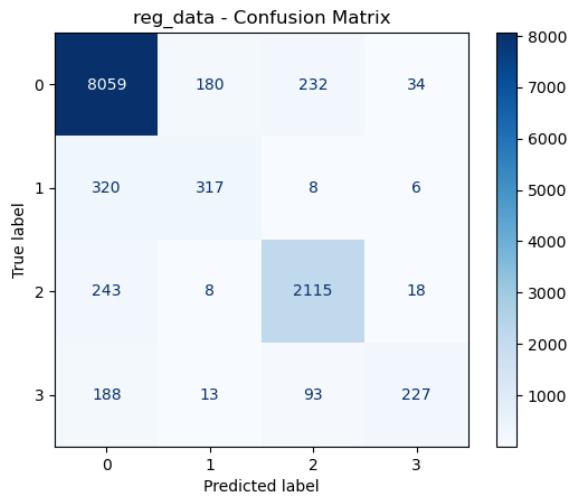


图 84: 混淆矩阵

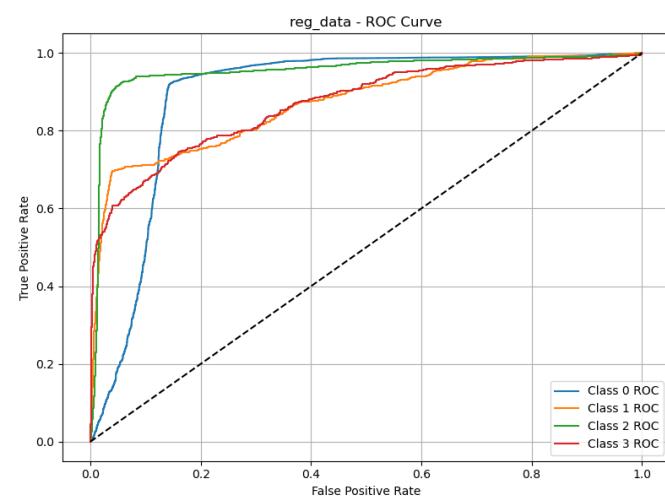


图 85: ROC 曲线

## 10 项特征提取数据

表 59: SVM 在 10 项特征提取数据训练下各项评估指标均值

Accuracy	Precision_macro	Recall_macro	F1_macro
0.8508	0.7249	0.5837	0.6298
AUC	Precision_weighted	Recall_weighted	F1_weighted
0.8411	0.8400	0.8508	0.8389

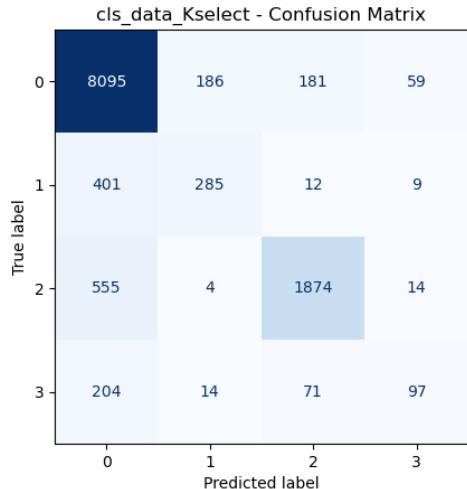


图 86: 混淆矩阵

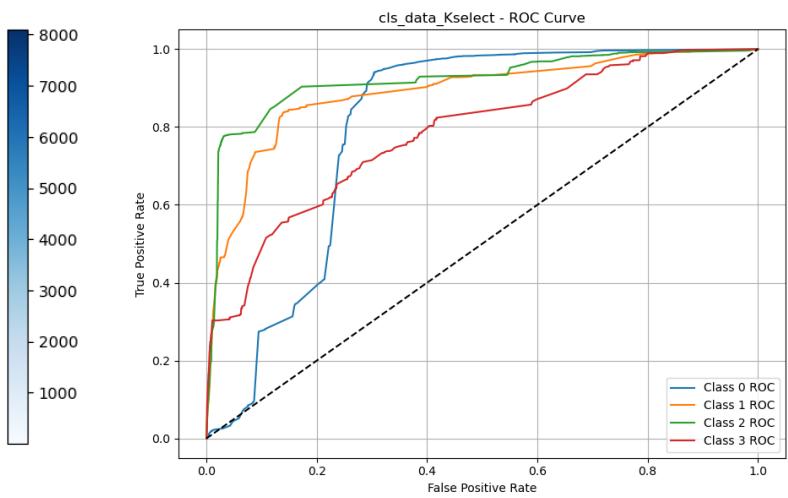


图 87: ROC 曲线

## 清洗后的 10 项特征提取数据

表 60: SVM 在清洗后的 10 项特征提取数据训练下各项评估指标均值

Accuracy	Precision_macro	Recall_macro	F1_macro
0.8824	0.4794	0.4230	0.4312
AUC	Precision_weighted	Recall_weighted	F1_weighted
0.7689	0.8452	0.8824	0.8580

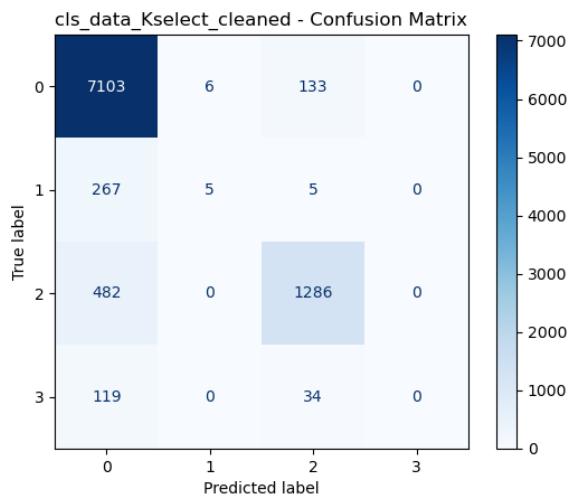


图 88: 混淆矩阵

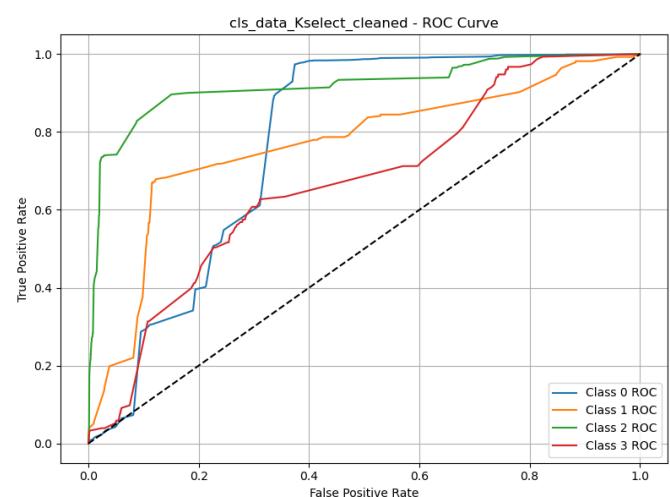


图 89: ROC 曲线