# 新生培训比赛报告

李俊江

## 1 数据可视化与分析

如图 1 所示，本数据集中的食材使用数量是非常不均衡的，有些食材的被使用数量非常少，甚至接近于 0，并且不同菜系的食材使用数量也是有较大差异的，偏好的食材各有差异。

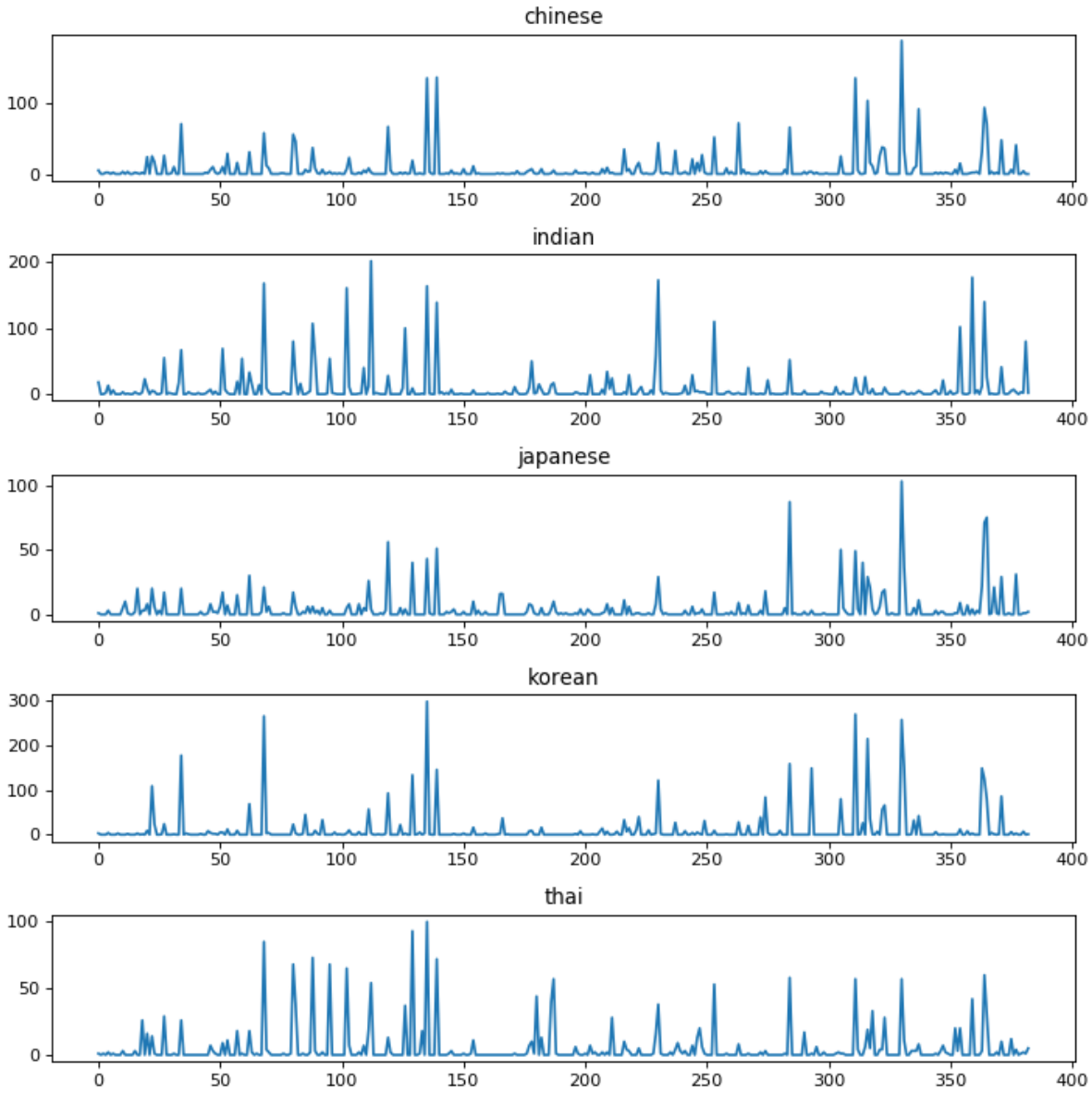



Figure 1: 各菜系的食材使用数量图

如图 2 所示，本数据集中菜系的数量从大到小依次为：korean、indian、chinese、japanese、thai。

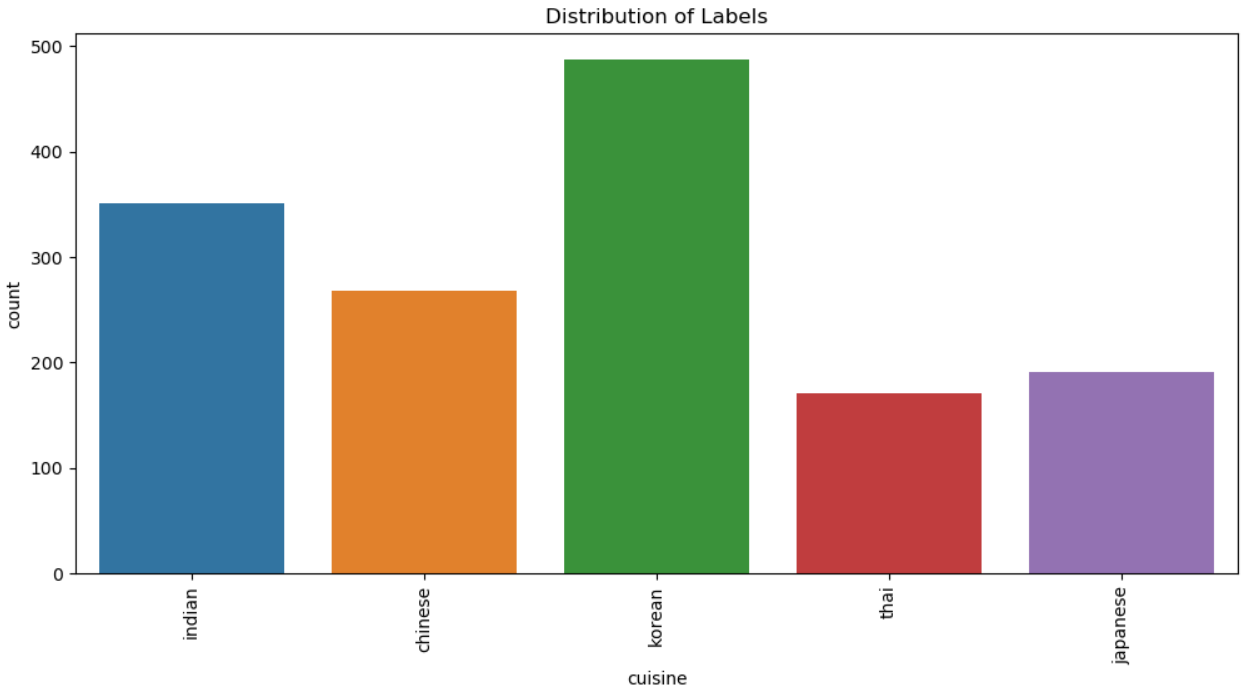



Figure 2: 各菜系的数量图

如图 3 所示，中国菜使用频率前十的食材依次是 soy sauce（酱油）、ginger（姜）、garlic（蒜）、scallion（葱）、sesame oil（芝麻油）、vegetable oil（植物油）、starch（淀粉）、pork（猪肉）、vinegar（醋）、black pepper（黑胡椒）。

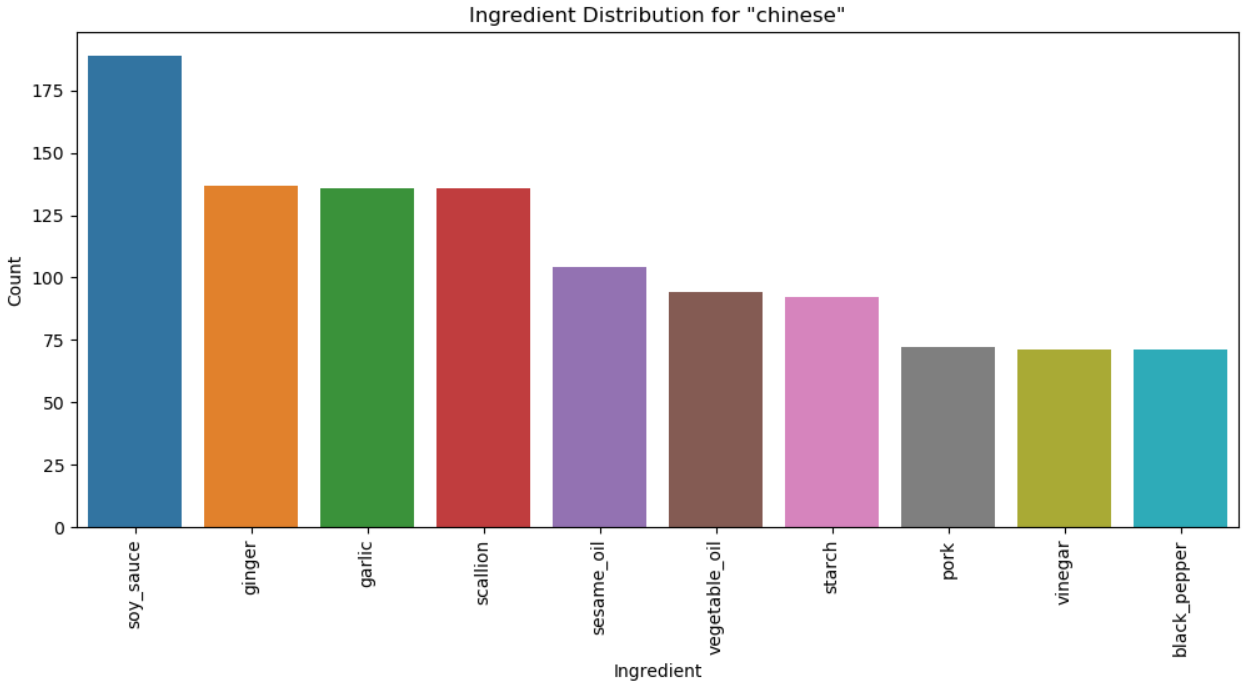



Figure 3: 中国菜的食材使用数量排名图

利用主成分分析（PCA），我们可以将特征降维到二维空间，并通过散点图展示标签在降维后的空间中的分布情况。

如图 4 所示，PC1 可以将菜系大致分为两类，左侧的 korean、japanese、chinese，右侧的 thai、indian。而 PC2 对菜系的区分度不够高，其中只有 japanese 位于下侧，其余菜系上下侧均有。

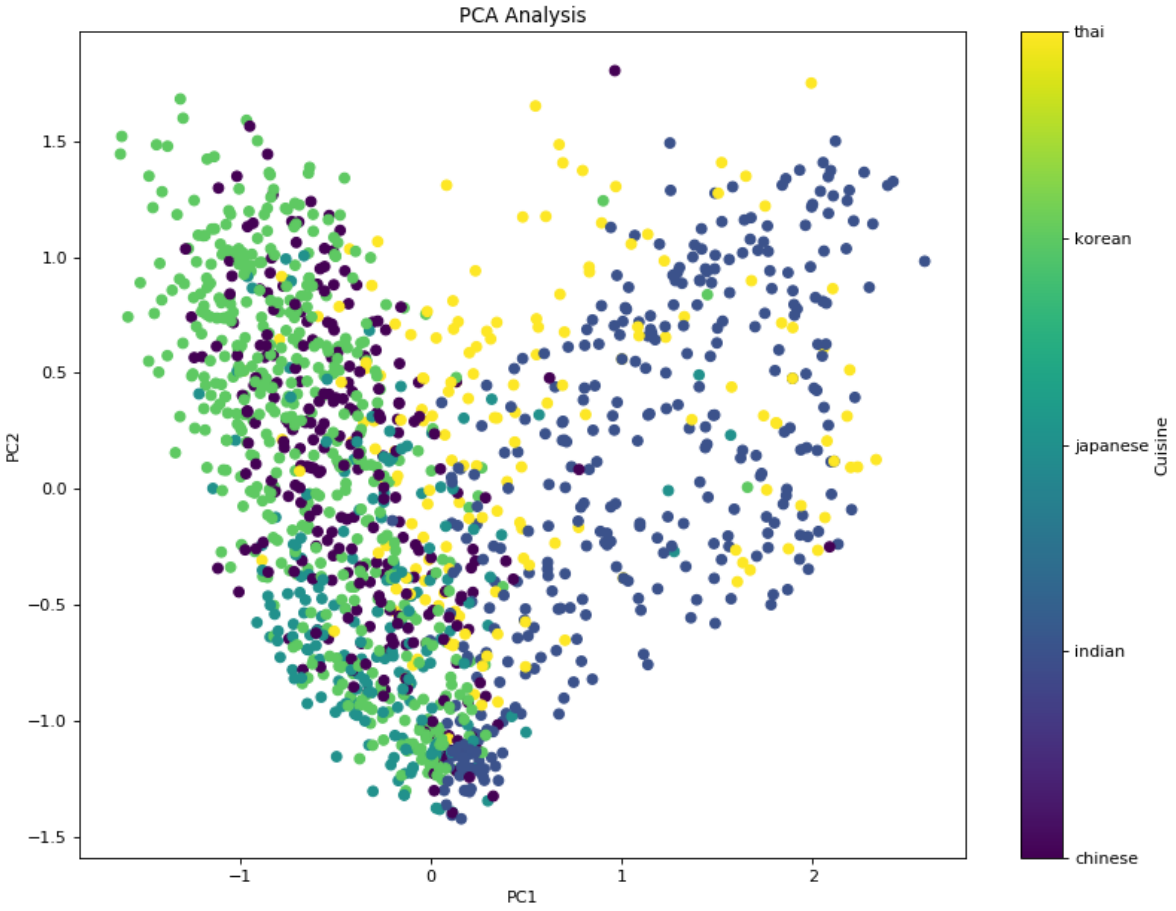



Figure 4: PCA 分析散点图

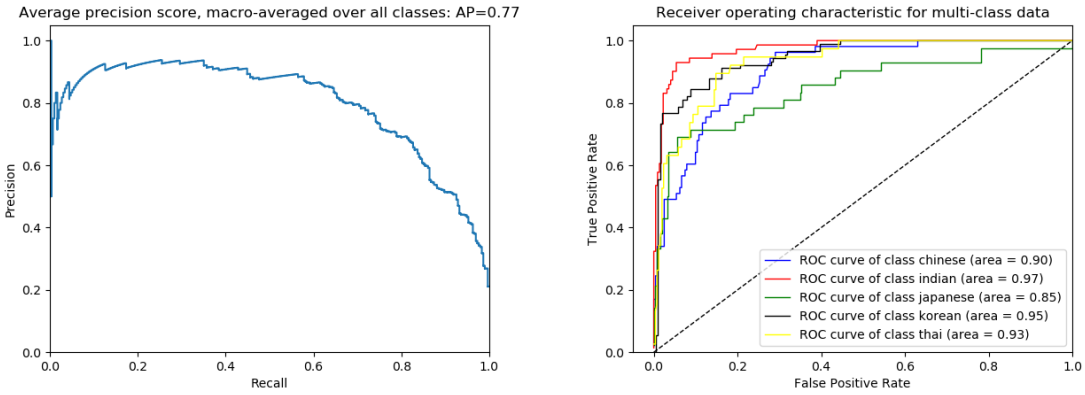
## 2 模型与算法

本实验分别实现了逻辑回归模型、集成模型和多层感知机，其实验结果如表 1 所示。可以看到一个神奇的现象，Log-Bagging 模型在训练数据集上得分最低，但其在公开数据集上得分最高，而 MLP 则相反。猜测是由于数据集太少，而特征太多，导致 MLP 模型的泛化性不够。

Table 1: 实验结果表

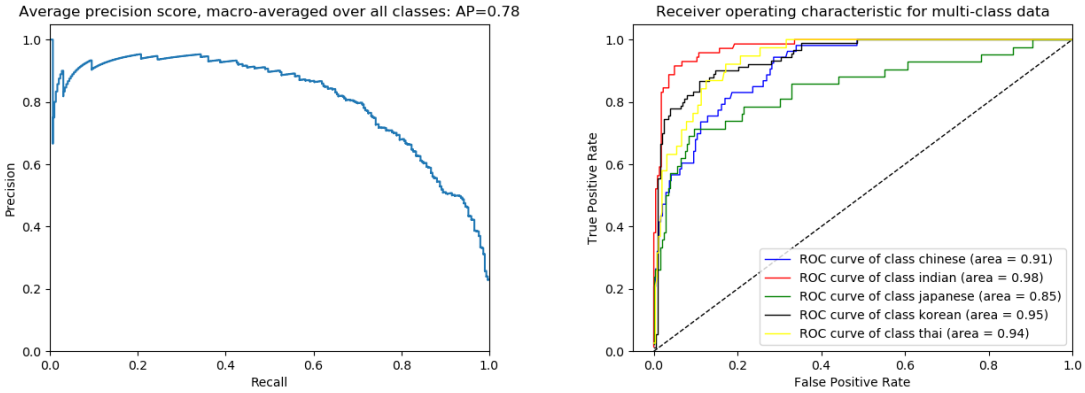
| Model | Train Score | Valid Score | Public Score |
|---|---|---|---|
| Log | 0.93 | 0.74 | 0.81 |
| Log-Bagging | 0.91 | 0.73 | 0.82 |
| MLP | 0.98 | 0.74 | 0.79 |

逻辑回归模型、集成模型和多层感知机的 PR 曲线和 ROC 曲线分别如图 5-7 所示。



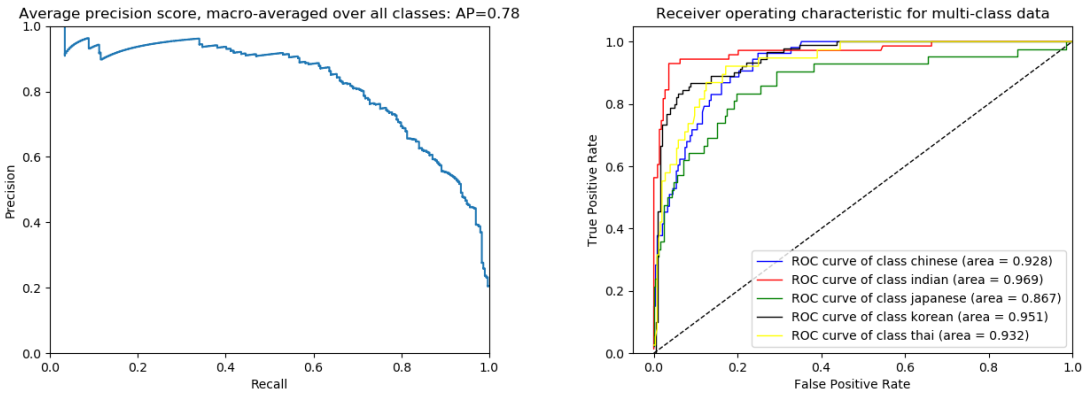

(a) PR 曲线　　(b) ROC 曲线

Figure 5: 逻辑回归模型的 PR 曲线和 ROC 曲线




(a) PR 曲线　　(b) ROC 曲线

Figure 6: 集成模型的 PR 曲线和 ROC 曲线




(a) PR 曲线　　(b) ROC 曲线

Figure 7: MLP 模型的 PR 曲线和 ROC 曲线

# 3 git 训练

由于是第一次使用 git 进行版本管理，所以在写代码的过程中忘记 commit 了，只在最后写完代码后 commit 了最终版。同时创建了共四个分支，main、model-SGD、model-integration、model-MLP，并在完成每个模型的训练预测且提交最终结果后，将对应分支合并到了 main 分支中。

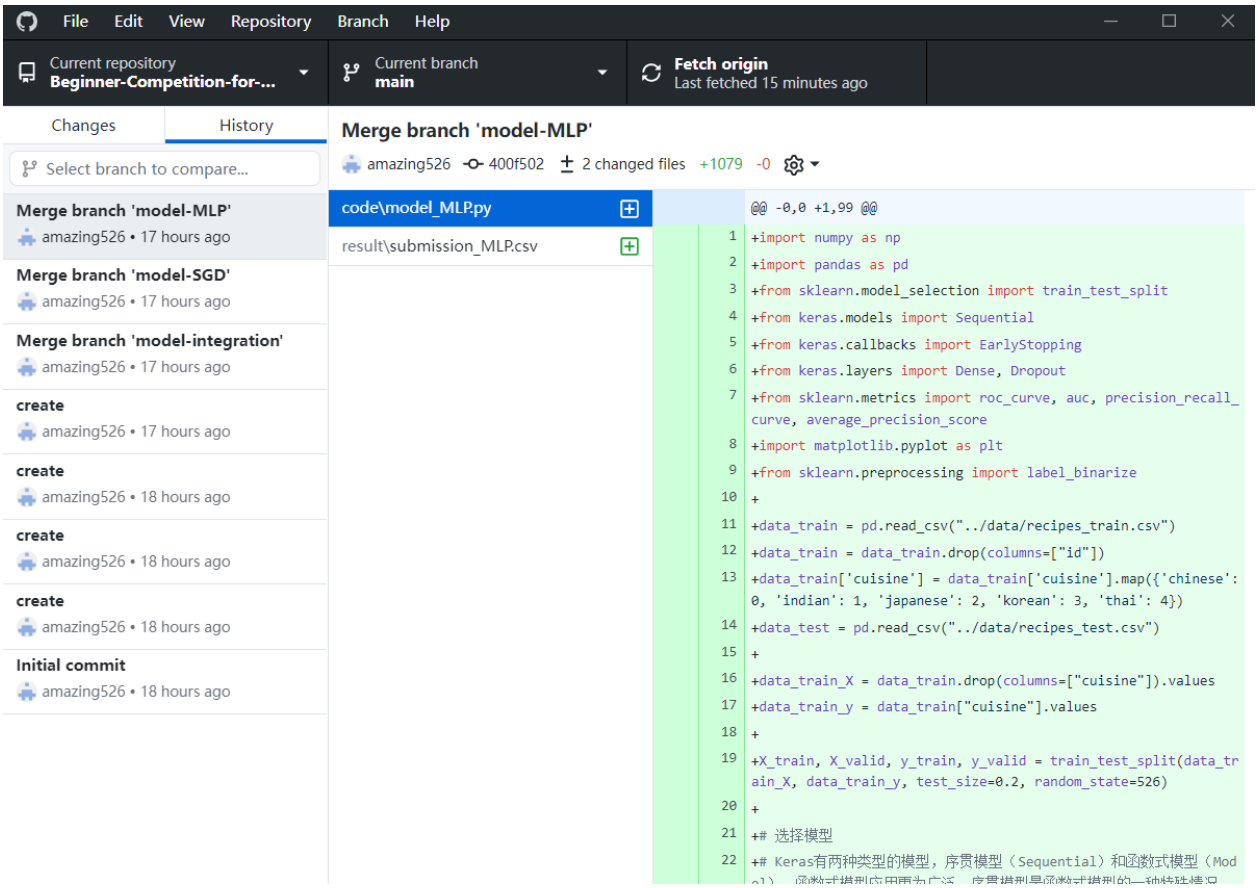



Figure 8: GitHub Desktop 软件的 git log 截图