**模块C：机器学习（40分）**

**竞赛题目：橙汁销售数据预测**

**背景介绍**

仓储式超市短期经营目标有两个：一是追求利润最大化，二是追求市场占有率最大化，即尽量扩大销售额，成功的经营应是在追求销售额增长的前提下追求利润增长。不论选择侧重于追求利润或是销售额增长的目标，它都必须通过商品的合理的配置陈列来实现。因此对商品的畅销程度有一定了解至关重要，畅销产品销售速度快，应该合理安排库存，得到及时补充。

**数据描述**

(数据集位于data/orange\_juice.csv需自行划分为test和train)数据由2个文件组成：train.csv，test.csv。其中train.csv为训练数据集，test.csv为测试数据集。训练数据集文件内含 845 个样本，每个样本包含 17 个特征和 1 个标签；测试数据集文件内含 212 个样本，每个样本包含 17 个 特征。

| **字段** | **说明** |
| --- | --- |
| Purchase | 0代表CH，1代表MM，指示客户是否购买了柑橘山或美汁橙汁 |
| WeekofPurchase | 购买周数 |
| StoreID | 商店ID |
| PriceCH | CH的成本价格 |
| PriceMM | MM的成本价格 |
| DiscCH | CH的价格折扣 |
| DiscMM | MM的价格折扣 |
| SpecialCH | CH的特殊指示符 |
| SpecialMM | MM的特殊指示符 |
| LoyalCH | 客户对CH的品牌忠诚度 |
| SalePriceMM | MM的销售价格 |
| SalePriceCH | CH的销售价格 |
| PriceDiff | MM的销售价格减去CH的销售价格 |
| Store7 | “否”和“是”级别的因子，指示是否在商店7中进行销售 |
| PctDiscMM | MM的折扣百分比 |
| PctDiscCH | CH的折扣百分比 |
| ListPriceDiff | MM的标签价格减去CH的标签价格 |
| STORE | 销售发生在5家可能的商店中的哪家 |

提示：其中Purchase字段为标签

**评分标准**

**目标**

本次数据为橙汁商品数据集，目标为训练一个分类模型，预测顾客更倾向于买哪种橙汁，并在测试数据集上尝试获得最优结果。

**评价指标**

accuracy\_score：sklearn.metrics.accuracy\_score

**环境介绍**

在本竞赛模块环境中，已自动启动了 jupyter notebook 环境。环境中已安装 Python 中多种数据分析和机器学习库，包括：

* Numpy
* Scipy
* Pandas
* Matplotlib
* Seaborn
* Scikit-learn
* Xgboost
* Lightgbm
* 其他

**任务列表**

1. 数据获取
2. 数据探查
3. 数据处理
4. 模型训练
5. 结果保存

**任务启动**

**Step1. 数据获取**

* 训练数据路径为 ../data/train.csv
* 测试数据路径为 ../data/test.csv

**Step2. 数据探查**

数据探查是对数据质量的检验。数据探查阶段为建模提供了指导，可以快速和分析数据中的异常数据，初步了解数据特征。

**Step3. 数据预处理与特征工程**

每个特征对预测结果可能会有影响，所以需要对数据进行处理，下面一些步骤是非必需进行的，以你自己的判断设计这个过程。

* 空值处理
* 重复值处理
* 特征缩放
* 特征提取
* 特征选择
* 特征组合
* ……

**Step4. 模型训练**

* 请自己选用合适的模型或改善此模型的参数以提高分数，包括但不限定以下模型：
  + sklearn.ensemble.ExtraTreesClassifier
  + sklearn.ensemble.RandomForestClassifier
  + sklearn.ensemble.AdaBoostClassifier
  + sklearn.ensemble.GradientBoostingClassifier
  + sklearn.gaussian\_process.GaussianProcessClassifier
  + sklearn.linear\_model.LogisticRegression
  + sklearn.linear\_model.RidgeClassifier
  + sklearn.linear\_model.SGDClassifier
  + sklearn.naive\_bayes.GaussianNB
  + sklearn.neighbors.KNeighborsClassifier
  + sklearn.neural\_network.MLPClassifier
  + sklearn.svm.SVC
  + sklearn.svm.NuSVC
  + sklearn.tree.DecisionTreeClassifier
  + sklearn.tree.ExtraTreeClassifier
  + xgboost.XGBClassifier
  + lightgbm.LGBMClassifier

**结果保存**

**任务结束**

恭喜完成本次比赛，祝您取得好成绩。