基于历史数据预测未来股价涨跌

本次竞赛的目标是基于沪深 300 指数成分股的历史股价数据，通过建立机器学习模型来预测未来股价涨跌幅最大和最小的股票。选手需通过构建模型、训练和调优，预测并输出给定数据后一天沪深 300 指数成分股的涨跌幅最大和最小各 10 支股票，以此进行排名。

一、 比赛数据

1. 训练数据（train.csv）

a) 数据时间范围：2015 年 4 月 20 日至 2025 年 4 月 20 日

b) 数据包含字段：股票代码、日期、开盘价、收盘价、最高价、最低价、成交量、成交额、换手率等。

数据示例：

股票代码	日期	开盘	收盘	最高	最低	成交量	成交额	振幅	涨跌额	换手率	涨跌幅
000001	2015-04-20	29.23	28.49	29.23	28.14	31308	89789343	3.74	-0.63	5.4	-2.16
000001	2015-04-21	28.41	28.86	28.9	27.36	33180	94291701	5.41	0.37	5.72	1.3
…	…	…	…	…	…	…	…	…	…	…	…

选手可以使用这个数据训练模型，预测未来的股票涨跌。

2. 推理数据（test.csv）

a) 数据时间范围：2015 年 4 月 20 日至 2025 年 4 月 25 日

b) 数据包含字段：股票代码、日期、开盘价、收盘价、最高价、最低价、成交量、成交额、换手率等。

数据示例：

股票代码	日期	开盘	收盘	最高	最低	成交量	成交额	振幅	涨跌额	换手率	涨跌幅
000001	2025-04-21	29.23	28.49	29.23	28.14	31308	89789343	3.74	-0.63	5.4	-2.16
000001	2025-04-22	28.41	28.86	28.9	27.36	33180	94291701	5.41	0.37	5.72	1.3
…	…	…	…	…	…	…	…	…	…	…	…

选手需基于此数据输出股市涨跌的预测。

3. 实际结果数据（check.csv）

a）数据时间范围：2025 年 4 月 28 日

b）数据包含字段：涨幅最大和最小股票的代码各 10 支（共 20 支）

$$涨幅(\%) = \frac{今天的收盘价 - 昨天的收盘价}{昨天的收盘价} \times 100$$

数据示例（涨幅数值从大到小序）：

涨幅最大股票代码	涨幅最小股票代码
000001	000003
000002	000004
…	…

二、 提交结果

选手的任务是基于 train.csv 训练模型，基于 test.csv 数据，输出预测结果 result.csv（UTF-8 编码，格式同 check.csv），并与 check.csv 比对，计算排名分数。

三、 评估标准

1. 计算 F1 分数：

- 精度（Precision）：对于前 10 只预测股票中，实际在前 10 名的股票的比例。
- 召回率（Recall）：实际前 10 只股票中被预测正确的比例。
- F1 分数的计算如下：

- 对于涨跌幅最大的 10 只股票：

$$\mathrm{F1}_{up} = \frac{2 \times \mathrm{Precision}_{up} \times \mathrm{Recall}_{up}}{\mathrm{Precision}_{up} + \mathrm{Recall}_{up}}$$

- 对于涨跌幅最小的 10 只股票：

$$\mathrm{F1}_{down} = \frac{2 \times \mathrm{Precision}_{down} \times \mathrm{Recall}_{down}}{\mathrm{Precision}_{down} + \mathrm{Recall}_{down}}$$

2. 排名相关性（Rank Correlation）：

- 排名相关性考虑预测股票在结果中的排序位置与实际结果排序的接近度。这里我们使用 Spearman 秩相关系数来衡量排名的一致性。通过比较实际与预测股票的顺序，计算其相关性。

- Spearman 秩相关系数公式：

$$\text{Spearman Rank Correlation} = 1 - \frac{6\sum d_i^2}{N(N^2-1)}$$

其中 d_i 为第 i 个预测股票与实际股票在排序中的排名差，最大记为 N，N 为股票的总数（这里取 10）。

- 排名相关性计算：

- 对于涨跌幅最大的 10 只股票：

$$\text{Rank Correlation}_{up} = \text{Spearman Rank Correlation for 涨跌幅最大股票}$$

- 对于涨跌幅最小的 10 只股票：

$$\text{Rank Correlation}_{down} = \text{Spearman Rank Correlation for 涨跌幅最小股票}$$

3. 最终得分：

$$\text{Final Score} = 0.2 \times \text{F1}_{up} + 0.2 \times \text{F1}_{down} + 0.3 \times \text{Rank Correlation}_{up} + 0.3 \times \text{Rank Correlation}_{down}$$

四、 其他说明

1. 本项比赛可以使用开源且可免费获取的数据集，但必须在提交结果中说明开源数据以及获取来源；

2. 可以使用开源预训练模型，该预训练模型需满足下列条件之一：

a) 使用非商业化公开数据集训练得到的预训练模型；

b) 已经在 2025 年 5 月 1 日前，在学术期刊、会议（不含 arxiv）、各大平台（如 Pytorch，Tensorflow，Github 等）发表的公开预训练模型；

3. 模型的可复现性以及创新性将会作为参考指标。
4. 线上赛C阶段，竞赛平台系统将更换数据集（格式不变）如下：
   a) 训练数据（train.csv）时间范围：2015年4月20至2025年7月18日
   b) 推理数据（test.csv）时间范围：2015年4月20日至2025年7月25日
   c) 最终实际结果数据（check.csv）包含7月28日股市涨跌幅最大和最小股票各10支的代码。