# 大数据挑战赛 基于历史数据预测未来股价涨跌

**一、赛题描述**

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五天后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，预测并输出五天后沪深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。

**二、比赛数据**

1. **训练数据（train.csv）**
   * 数据时间范围：2015年4月20日至2025年4月20日
   * 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。

1. **测试数据（test.csv）**
   * 数据时间范围：2015年4月21日至2025年4月25日
   * 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。

1. **实际结果数据（check.csv）**
   * 数据时间范围：2025年4月28日
   * 数据包含字段：涨跌幅最大和最小股票的代码各10支（共20支）

数据示例（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 |
| ... | ... |

选手的任务是基于test.csv数据，输出预测结果result.csv（格式同check.csv），并与check.csv比对，计算排名分数。

**三、比赛任务与流程**

1. **任务**
   * 选手需开发一个预测模型，根据给定的历史股价数据（test.csv），预测未来1天股价涨跌幅最大和最小的各10只股票，并输出股票代码（result.csv）。
2. **输入与输出**
   * **训练：**
     + train.csv：历史股价数据。
   * **输入：**
     + test.csv：5天的股价数据。
   * **输出：**
     + result.csv：根据模型预测结果输出涨跌幅最大和最小各10支股票的代码。
3. **评分规则**
   * 每日提交的预测结果（result.csv）将与实际结果（check.csv）比对，计算准确度并排名。
   * 详见详细评分公式。
4. **提交要求**
   * **预测模型**：选手需提供完整的代码（python）、训练数据和训练模型。
   * **代码规范**： 选手提交的文件夹顺序如下：

|--app

|--code（储存运行代码，docker内）

|--src

|--featurework.py

|--test.py

|--train.py

|--data（储存数据，在docker-compose.yml中进行挂载）

|--test.csv

|--train.csv

|--model（储存训练模型，docker内）

|--output（储存计算结果，在docker-compose.yml中进行挂载）

|--top\_10\_max\_target.csv

|--top\_10\_min\_target.csv

|--temp（储存中间结果，在docker-compose.yml中进行挂载）

|--init.sh（docker内）

|--train.sh（docker内）

|--test.sh（docker内）

选手需要按照上面所示将文件进行存储，在docker中储存code，model，init.sh，train.sh，test.sh等文件，将docker镜像命名为bdc2025，并导出为队伍名称.tar进行提交，我们会使用docker load -i 队伍名称.tar命令对docker镜像进行加载，并使用下发的docker-compose.yml文件运行。

* + **性能要求**：确保代码可以在指定机器（i7-13650H，16GB内存，4060独显8GB显存，500GB存储）上运行并生成结果。
    - **预测时间**：选手的模型预测时间不得超过5分钟。
    - **训练时间**：模型训练时间不得超过8小时。

1. **模型与资源**
   * 选手可使用任何公开的免费数据资源和机器学习模型资源；
   * 需要明确说明；
   * 训练和预测时不得联网。
2. **时间安排**
   * **报名及竞赛阶段**：2025年5月5日 - 2025年7月18日
     + 每日提交预测结果，并实时更新榜单。
   * **最终提交**：2025年7月20日23:59之前
     + 最终代码和模型提交；
     + 所使用的所有外部数据和模型资源。
3. **最终测试数据（test.csv）**
   * 竞赛结束后5天的数据，时间范围：2025年7月21日至2025年7月25日
   * 格式与前述test.csv相同。
4. **最终实际结果数据（check.csv）**
   * 数据格式与前述check.csv相同，包含7月28日股市涨跌幅最大和最小股票各10支的代码。
5. **生成排名**：2025年7月29日
   * 根据最终预测结果生成排行榜。
6. 代码复审和重现
   * 针对排名前30的进行代码复审和重现
   * 重现时，使用训练数据（train.csv）时间范围：2015年4月20至2025年7月18日

**四、详细评分公式**

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：

- 对于涨跌幅最小的10只股票：

1. **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：

其中 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：

- 对于涨跌幅最小的10只股票：

1. **最终得分：**