中国高校计算机大赛

2025 中国高校计算机大赛 —— 大数据挑战赛

通　知

2016 年，教育部高等学校计算机类专业教学指导委员会、教育部高等学校软件工程专业教学指导委员会、教育部高等学校大学计算机课程教学指导委员会、全国高等学校计算机教育研究会联合创办了“中国高校计算机大赛”（China Collegiate Computing Contest，简称 C4），目前“中国高校计算机大赛”继续由全国高等学校计算机教育研究会主办。大数据挑战赛是其中的一项重要赛事，在 2018-2024 年均入选全国普通高校学科竞赛排行榜，获得社会各界的高度关注和广泛好评。

2025 中国高校计算机大赛——大数据挑战赛（以下简称“大赛”）由清华大学、大数据系统软件国家工程研究中心联合举办。由上海和今信息科技有限公司提供竞赛平台支持。大赛是以实际数据为基础、面向全球开放的高端算法竞赛。大赛旨在通过竞技的方式，提升人们对数据分析与处理的算法研究与技术应用能力，探索大数据的核心科学与技术问题，尝试创新大数据技术，推动大数据的产学研用。

本次大赛面向全球开放，不限年龄国籍，高等院校在校学生（包括高职高专、本科生、研究生）以及科研机构和企业从业人员均可报名参赛。参赛队伍根据赛题要求设计相应的算法进行数据分析和处理，比赛结果按照指定的评价指标使用在线评测数据进行评测和排名，得分最优者获胜。

请各学校积极配合，按照通知和大赛章程做好宣传和组织工作，为在校生和毕业生参与竞赛提供必要的条件和支持。

竞赛详情见附件（2025 大数据挑战赛竞赛规程）。

全国高等学校计算机教育研究会

2025 年 5 月

2025 中国高校计算机大赛——大数据挑战赛
竞赛规程

2016 年，教育部高等学校计算机类专业教学指导委员会、教育部高等学校软件工程专业教学指导委员会、教育部高等学校大学计算机课程教学指导委员会、全国高等学校计算机教育研究会联合创办了“中国高校计算机大赛”（China Collegiate Computing Contest，简称 C4），目前“中国高校计算机大赛”继续由全国高等学校计算机教育研究会主办。大数据挑战赛是其中的一项重要赛事，在 2018-2024 年期间均入选全国普通高校学科竞赛排行榜，是国内高校 A 类赛事，获得社会各界的高度关注和广泛好评。

2025 中国高校计算机大赛——大数据挑战赛（以下简称“大赛”）由清华大学、大数据系统软件国家工程研究中心主办。由上海和今信息科技有限公司提供竞赛平台支持。大赛是以实际数据为基础、面向全球开放的高端算法竞赛。大赛旨在通过竞技的方式，提升人们对数据分析与处理的算法研究与技术应用能力，探索大数据的核心科学与技术问题，尝试创新大数据技术，推动大数据的产学研用。

本次大赛聚焦于时间序列数据的建模与预测，通过构建基于真实金融市场数据的任务场景，旨在推动前沿算法在实际复杂环境中的落地应用。时间序列数据广泛存在于金融、交通、能源、医疗等领域，具有强烈的时序依赖性和动态变化特征。股价作为典型的时间序列对象，表现出高波动性、高频率、强非线性和多因素驱动等复杂特性，对建模技术提出了严峻挑战。

本次竞赛选择中国 A 股市场的股价数据作为研究对象，是基于其高度代表性和数据质量的综合考量。A 股市场是中国资本市场的核心组成部分，包含上千家上市公司，涵盖多个行业和市值层级，拥有丰富的历史数据与活跃的交易行为。其股价受宏观经济政策、行业发展、企业基本面、市场情绪等多重因素影响，为时间序列预测模型的特征提取、机制建模、异常识别与动态调整提供了良好的实践平台。同时，A 股股价数据在时间维度上既具备微观的高频波动，也蕴含中长期的趋势变化，有助于推动参赛者设计多尺度、分层次的预测方法，提升模型的综合表现。

从更宏观的视角来看，A 股市场在中国经济发展与全球资本格局中的地位日益重要。近年来，随着注册制改革的全面推进、新兴科技企业的持续上市、以及资本市场对外开放步伐的加快，A 股正逐步从融资市场向资源配置市场转变，承担起更为核心的经济调节与资源优化职能。在国家大力发展数字经济、推动金融科技融合创新的背景下，围绕 A 股市场构建智能化分析与预测工具，不仅能够服务于投资决策与风险控制，也具有提升金融体系智能化水平、促进技术成果转化的现实意义。

一、赛题设置

本次大赛的题目是“基于历史数据预测未来股价涨跌”，具体说明详见附件“赛题描述”。其目标是基于沪深 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个交易日沪深 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、参赛对象

本次大赛面向全球开放，不限年龄国籍，高等院校在校学生（包括高职高专、本科、研究生）以及科研机构和企业从业人员均可参赛。具体要求如下：

- 可以自由组队参赛，具体组队要求见后文相关说明；
- 参赛选手应保证报名信息准确有效，如队伍中的选手信息不符合要求，组委会有权取消整个队伍的参赛资格及奖励。

三、赛制说明

本次大赛分为报名&组队、线上赛和决赛等三个阶段，其中线上赛均由参赛队伍下载数据在本地进行算法设计和调试，并通过大赛报名官网提交结果文件及模型代码；决赛要求参赛者进行现场演示和答辩。

1. 报名&组队（5 月 20 日 – 7 月 15 日）

参赛选手须在竞赛平台报名并且组队参赛（即使单人参赛也要组建单人队伍），大赛不收取任何报名费用。大赛报名系统开放时间为北京时间 2025 年 5 月 20 日 10:00，截止时间为北京时间 2025 年 7 月 15 日中午 12:00。

- 报名方式：登录竞赛平台，完成个人信息注册，即可报名参赛；
- 每个选手可单人成队或 2-3 人组队参赛；
- 参赛队伍（包括队长及全体队伍成员）需要在竞赛平台完成实名认证，未完成认证的队伍将无法参加正式比赛。

大赛官方渠道主要包括：

- 大赛官网：https://nercbds.tsinghua.edu.cn/bdc.html
- 竞赛平台：https://www.heywhale.com/home/
- 大赛邮箱：data@tsinghua.edu.cn
- 大赛 QQ 群：762146461 / 901317172

报名截止之后，不再允许添加或更改任何队伍成员。如有中途退出情况，只允许在参赛队伍内部更换队长或删除队员。参赛队伍须应在决赛开始前向大赛组委会提交成员更换申请，由参赛队伍全部成员亲笔签名，经由大赛组委会审核后变更生效。

2. 线上赛（5 月 20 日 – 7 月 20 日）

参赛队伍可从竞赛平台下载数据，在本地进行算法调试，并在线提交结果及模型代码。若参赛队伍在一天内多次提交结果，新结果版本将覆盖旧版本。

线上赛 A 阶段：5 月 20 日 10:00 – 7 月 18 日 20:00，每个参赛队伍每天可以进行 2 次结果提交，系统立即进行评测并返回成绩。排行榜实时进行更新，将选择参赛队伍在本阶段的历史最优成绩进行排名展示。请确保结果可复现。

线上赛 B 阶段：7 月 19 日 – 7 月 20 日 23:59，每个参赛队伍提交整理好的模型代码，要求详见“代码规范”文档。

线上赛 C 阶段：7 月 28 日 –，系统将在 7 月 28 日 20:00 更换训练数据和推理数据，并运行选手模型代码获得结果文件进行计算排名展示。

线上赛结束后，排名前 70 名的参赛队伍以及排名在 71-110 之间前 30 支学生队伍将进行代码审核。组委会将审核并剔除没有机器学习算法贡献的队伍，并取消存在违反比赛规定队伍的比赛资格，空缺名额不再替补。所有通过审核的队伍将获得线上赛名次证书。

3. 决赛（8 月中下旬）

决赛将以现场答辩会的形式进行，具体要求和安排另行通知。受邀参加决赛的选手在决赛期间的食宿由大赛组委会负责，其他费用自理。

四、奖项设置

大赛的奖金池总额为 5 万元人民币，所有奖金均为税前金额。

1. 线上赛奖项（以大赛官网线上赛最终排行榜为准）

线上赛通过代码审核的 100 支队伍将颁发线上赛名次证书。

2. 决赛奖项（以大赛官网决赛结果为准）

奖励对象	数量	奖励办法
决赛第 1 名队伍	1	奖金 2 万元，决赛名次证书
决赛第 2 名队伍	1	奖金 1 万元，决赛名次证书
决赛第 3 名队伍	1	奖金 0.8 万元，决赛名次证书
决赛第 4-6 名队伍	3	奖金 0.4 万元，决赛名次证书

3. 在校学生队伍奖项

在校学生队伍要求所有参赛队员必须全部为在校学生，如果队伍中有一名在职人员，则整个队伍视为在职人员队伍。其中中国大陆在校学生提供学信网的教育部学籍在线验证报告编号进行身份验证，其余学生提供相关在读证明进行身份验证，在校学籍以 2025 年 5 月 30 日为准。

此奖项仅颁发给在校学生队伍，要求队伍通过代码审核，并根据在校学生队伍成绩的单独排名结果进行颁发。

奖项名称	数量	对象
全国一等奖	5	单独排名第 1-5 名
全国二等奖	10	单独排名第 6-15 名
全国三等奖	15	单独排名第 16-30 名

五、违规处理

参赛者应本着诚实、公平的态度参加比赛，如在以下情况出现违规，大赛组织委员会（简称“组委会”）有权取消参赛者所在队伍的参赛资格，情节严重者将通报参赛者所在单位并追究其违法责任。

1. 账号使用：参赛者所用的账号必须是使用本人信息注册的，并有义务保证账号所有信息的真实性和有效性，且账号仅限于参赛者本人使用；参赛者禁止使用多账号参赛，同一参赛者不可使用多个账号进行提交、刷分操作；如根据判断认为参赛账号存在异常或违背正常使用条例，组委会可以单方面暂停或终止该账号登录大赛平台。

2. 比赛成果：
    - 严禁参赛队伍之间相互抄袭。如不同参赛队伍提交结果高度相似，经判定存在抄袭行为的，组委会将取消相关参赛队伍的参赛资格，相关参赛成绩无效。
    - 参赛者应保证其在比赛过程中所产出的所有成果未侵犯任何第三方的知识产权、商业秘密及其他合法权益。如第三方因为参赛者侵权行为提出索赔、诉讼等，参赛者应承担由此产生的全部责任及损失。
    - 如大赛主办方及其关联公司有意取得参赛者在本次大赛中独立开发的依约定享有完整知识产权的研究成果，参赛者同意大赛主办方及其关联公司在同等条件下享有优先受让权，相关转让事宜由双方另行协商确定。

3. 数据使用：对于大赛提供的数据（数据集），参赛者须仅在比赛场景下使用，并应妥善保存已下载的数据（数据集），避免泄露；在完成比赛使用后应及时销毁已下载数据（数据集）；如使用比赛之外的任何数据应获得组委会许可。对于不提供下载的比赛数据，参赛者不得以任何形式擅自复制、下载或获取。参赛者如发现任何出现数据未授权访问的可能，应立即通知组委会并积极提供相关信息。如参赛者泄露已下载的数据（数据集），或未及时销毁已下载的数据（数据集）导致已下载的数据（数据集）泄露，参赛者应承担由此产生的全部责任及损失。

4. 代码分享：在大赛举办期间，未经组委会同意，参赛者禁止公开分享与赛事相关的数据、模型和代码；大赛结束之后，参赛者可以在拥有模型和代码的知识产权的情况下自行选择公开分享，但需要确保此类公开共享不会侵犯任何第三方的知识产权、商业秘密及其他合法权益。

5. 参赛者若在参赛过程中发现相关规则漏洞或技术漏洞，有义务及时告知组委会相关漏洞的信息，组委会将对提供相关信息的参赛者表示感谢；若参赛者利用相关漏洞进行参赛，经判断查证后，成绩将会被判断为无效成绩。

六、申诉与仲裁

1. 参赛团队或选手对不符合大赛规定的设备、工具和软件，有失公正的评判和奖励以及工作人员的违规行为等，均可向大赛组委会提出申诉。组委会负责受理比赛中提出的申诉并进行调解仲裁，以保证大赛的顺利进行和大赛结果的公平公正。组委会作出的仲裁结果为终局决定。
2. 申诉报告应明确申诉内容，指定一名成员作为联系人，通过大赛邮箱以邮件发送，否则申诉将不予受理。
3. 组委会将在收到申诉之日起 5 个工作日之内受理，并认真核查和处理。

七、其他

1. 为了确保整个大赛顺利、公正地进行，以及保证参赛选手的合法权益，参赛选手报名时应阅读和确认大赛官网上的《参赛协议》，自觉遵守协议规定。
2. 在大赛举办过程中，竞赛规程可能会有少量的变更和调整，大赛组委会将本着公平、公正、公开的原则在大赛官网公告，所有内容均以大赛官网为准。

“中国高校计算机大赛——大数据挑战赛”组织委员会

2025 年 5 月

附件：赛题描述—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）**

- 数据时间范围：2015 年 4 月 20 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。

3. **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）

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

数据示例（涨跌幅数值从大到小序）：

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。