A Accessibility, accountability and Maintenance

FinRL-Meta is an open-source project on GitHub, with MIT License. Codes, market environments, benchmarks and documentations are available on the GitHub repository: https://github.com/AI4Finance-Foundation/FinRL-Meta. FinRL-Meta has been actively maintained by FinRL community that has over 10K members at the moment. On GitHub, we keep updating our codes, merging pull requests, and fix bugs and issues. We welcome contributions from community members, researchers and quant traders.

B Dataset Documentation and Usages

We organize our contents according to the suggested template of datasheets for datasets.

B.1 Motivation

• For what purpose was the dataset created?
  As data refreshing minute-to-millisecond, finance is a particularly difficult playground for deep reinforcement learning.
  In academia, scholars use financial big data to obtain more complex and precise understanding of markets and economics. While industries use financial big data to refine their analytical strategies and strengthen their prediction models. To serve the rapidly growing FinRL community, we creates FinRL-Meta that provides data accessing from different sources and build the data to RL environments. We aim to provide dynamical RL environments that are manageable by users.

• Who created the dataset?
  FinRL-Meta is an open-source project created by the FinRL community. Contents of FinRL-Meta are contributed by the authors of this paper and will be maintained by members of FinRL community.

• Who funded the creation of the dataset?
  AI4Finance Foundation, a non-profit open-source community that shares AI tools for finance, funded our project.

B.2 Composition

• What do the instances that comprise the dataset represent?
  Instances of FinRL-Meta are financial data includes: stocks, securities, cryptocurrencies, etc. FinRL-Meta provides hundreds of market environments through an automatic pipeline that collects dynamic datasets from real-world markets and processes them into standard gym style market environments. FinRL-Meta also benchmarks popular papers as stepping stones for users to design new trading strategies.

• How many instances are there in total?
  FinRL-Meta does not store data directly. Instead, we provide codes for data accessing, data cleaning, feature engineering, and building into RL environments. Table B.1 provides the supported data sources of FinRL-Meta.
  At the moment, there are hundreds of market environments, tens of tutorials and demos, and several benchmarks provided.

• Does the dataset contain all possible instances or is it a sample of instances from a larger set?
  With our provided codes, users could fetch data from the data source by properly specifying the starting date, ending date, time granularity, asset set, attributes, etc.

• What data does each instance consist of?
  Now there are several types of financial data, as shown in Table B.1.

---

Table 3: Supported data sources. OHLCV means open, high, low, and close prices; volume data.

- Stocks
- Cryptocurrencies
- Securities
- ETFs
- Exchange rate:

- Is there a label or target associated with each instance?
  No. There is not label or preset target for each instance. But users can use our benchmarks as baselines.

- Is any information missing from individual instances?
  Yes. In several data sources, there are missing values and we provided standard preprocessing methods.

- Are relationships between individual instances made explicit?
  Yes. An instance is a sample set of the market of interest.

- Are there recommended data splits?
  We recommend users to follow our training-testing-training pipeline, as shown in Fig. 8. Users can flexibly choose their preferred settings, e.g., in stock trading task, our demo access Yahoo! Finance database and use data from 01/01/2009 to 06/30/2020 for training and data from 07/01/2020 to 05/31/2022 for backtesting.

- Are there any errors, sources of noise, or redundancies in the dataset?
  For the raw data fetched from different sources, there are noise and outliers. We provide codes to process the data and built them into standard RL gym environment.
Figure 8: Overview of FinRL-Meta training-testing-trading pipeline.

- **Is the dataset self-contained, or does it link to or otherwise rely on external resources?**
  It is linked to external resources. As shown in Table B.1, FinRL-Meta fetch data from data sources to build gym environments.

- **Does the dataset contain data that might be considered confidential?**
  No. All our data are from publicly available data sources.

- **Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?**
  No. All our data are numerical.

### B.3 Collection Process

- **How was the data associated with each instance acquired?**
  FinRL-Meta fetch data from data sources, as shown in Table B.1.

- **What mechanisms or procedures were used to collect the data?**
  FinRL-Meta provides dynamic market environments that are built according to users’ settings. To achieve this, we provide software APIs to fetch data from different data sources. Note that some data sources require accounts and passwords or have limitations on number or frequency of requests.

- **If the dataset is a sample from a larger set, what was the sampling strategy?**
  It is dynamic, depending on users’ settings, such as the starting date, ending date, time granularity, asset set, attributes, etc.

- **Who was involved in the data collection process and how were they compensated?**
  Our codes collect publicly available market data, which is free.

- **Over what timeframe was the data collected?**
  It is not applicable because the environments are created dynamically by running the codes to fetch data in real time.

- **Were any ethical review processes conducted?**
  No?

### B.4 Preprocessing/cleaning/labeling

- **Was any preprocessing/cleaning/labeling of the data done?**
  Yes. For the raw data fetched from different sources, there are noise and outliers. We provide codes to process the data and built them into standard RL gym environment.

- **Was the raw data saved in addition to the preprocessed/cleaned/labeled data**
  The raw data are held by different data sources (data providers).

- **Is the software that was used to preprocess/clean/label the data available?**
  Yes. We use our own codes to do cleaning and preprocessing.
B.5 Uses

• Has the dataset been used for any tasks already?
  Yes. Thousands of FinRL community members use FinRL-Meta for learning and research purpose. Demos and tutorials are mentioned in Section 3.

• Is there a repository that links to any or all papers or systems that use the dataset?
  1. Research papers that used FinRL-Meta are listed here: https://github.com/AI4Finance-Foundation/FinRL/blob/master/tutorials/FinRL_papers.md
  2. The following three repositories have incorporated FinRL-meta:
     – FinRL-meta corresponding to the market layer of FinRL (5.6K stars): https://github.com/AI4Finance-Foundation/FinRL
     – ElegantRL (2.1K stars) supports FinRL-Meta: https://github.com/AI4Finance-Foundation/ElegantRL
     – FinRL-Podracer: https://github.com/AI4Finance-Foundation/FinRL_Podracer

• What (other) tasks could the dataset be used for?
  Besides the current tasks (tutorial, demo and benchmarks), FinRL-Meta will be useful for the following tasks:
  – Curriculum learning for agents: Based on FinRL-meta (a universe of market environments, say $\geq 100$), one is able to construct an environment by sampling data samples from multiple market datasets, similar to XLand [39]. In this way, one can apply the curriculum learning method [39] to train a generally capable agent for several financial tasks.
  – To improve the performance for the large-scale markets, we are exploiting GPU-based massive parallel simulation such as Isaac Gym [29].
  – It will be interesting to explore the evolutionary perspectives [17, 33, 20, 24] to simulate the markets. We believe that FinRL-Meta will provide insights into complex market phenomena and offer guidance for financial regulations.

• Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?
  We believe that FinRL-Meta will not encounter usage limit. Our data are fetched from different sources in real time when running the codes. However, there may be one or two out of $\geq 30$ data sources (in Table B.1) change data access rules that may impact future use. So please refer to the rules and accessibility of certain data source when using.

• Are there tasks for which the dataset should not be used?
  No. Since there are no ethical problem for FinRL-Meta, users could use FinRL-Meta in any task as long as it does not violate laws.

  Disclaimer: Nothing herein is financial advice, and NOT a recommendation to trade real money. Please use common sense and always first consult a professional before trading or investing.

B.6 Distribution

• Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created?
  No. It will always be held on GitHub under MIT license, for educational and research purpose.

• How will the dataset be distributed?
  Our codes and existing environments are available on GitHub FinRL-Meta repository https://github.com/AI4Finance-Foundation/FinRL-Meta.

• When will the dataset be distributed?
  FinRL-Meta is publicly available since February 14th, 2021.
• Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?
  FinRL-Meta is distributed under MIT License, for educational and research purpose.

• Have any third parties imposed IP-based or other restrictions on the data associated with the instances?
  No.

• Do any export controls or other regulatory restrictions apply to the dataset or to individual instances?
  No. Our data are fetched from different sources in real time. However, there may be one or two out of ≥ 30 data sources (in Table B.1) change data access rules that may impact future use. So please refer to the rules and accessibility of certain data source when using.

B.7 Maintenance

• Who will be supporting/hosting/maintaining the dataset?
  FinRL-Meta has been actively maintained by FinRL community (including the authors of this paper) that has over 10K members at the moment. We are still actively updating market environments, to serve the rapidly growing FinRL community.

• How can the owner/curator/manager of the dataset be contacted?
  To contact the main developers, we encourage users join our Slack channel: https://join.slack.com/t/ai4financeworkspace/shared_invite/zt-v56/0l1jm-d21g119fH2l1j1r2prFY0kg or our mailing list: https://groups.google.com/u/1/g/ai4finance.

• Is there an erratum?
  Users can use GitHub to report issues/bugs, and use Slack channel or mailing list to discuss solutions. FinRL community is actively improving the codes, say extracting technical indicators, evaluating feature importance, quantifying the probability of backtesting overfitting, etc.

• Will the dataset be updated?
  Yes, we are actively updating FinRL-Meta’s codes and data sources. Users could get information and the newly updated version through our GitHub repository, or join the mailing list: https://groups.google.com/u/1/g/ai4finance.

• If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances
  The data of FinRL-Meta do not relate to people.

• Will older versions of the dataset continue to be supported/hosted/maintained?
  Yes. All versions can be found on our GitHub repository.

• If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?
  We maintain FinRL-Meta on GitHub. Users can use GitHub to report issues/bugs, and use Slack channel or mailing list to discuss solutions. We welcome community members to submit pull requests through GitHub.
C Benchmarks

We have reproduced experiments in several papers as benchmarks. Our codes are publicly available for research purpose. Users can utilize them as stepping stones for designing trading strategies.

The codes of these benchmarks can be found at https://github.com/AI4Finance-Foundation/FinRL/tree/master/tutorials. Each notebook gives a full tutorial on how to reproduce the experiment of the paper.

C.1 Stock Trading

We access Yahoo! Finance database and select the 30 constituent stocks (accessed at 07/01/2020) in Dow Jones Industrial Average (DJIA). We use data from 01/01/2009 to 06/30/2020 for training and data from 07/01/2020 to 05/31/2022 for testing. We use technical indicators in our state space, e.g., Moving Average Convergence Divergence (MACD), Relative Strength Index (RSI), Commodity Channel Index (CCI), Average Directional Index (ADX), etc.

Code available at: https://github.com/AI4Finance-Foundation/FinRL/blob/master/tutorials/1-Introduction/FinRL_StockTrading_Fundamental.ipynb

C.2 Liquidation Analysis

We build a simulated environment of stock prices according to the Almgren and Chriss model. Then we implement the multi-agent DRL algorithms for both competing and cooperative liquidation strategies. This benchmark demonstrates the trade execution task using deep reinforcement learning algorithms. When trading, traders want to minimize the expected trading cost, which is also called implementation shortfall.

Code available at: https://github.com/AI4Finance-Foundation/FinRL-Meta/tree/master/tutorials/2-Advance/execution_optimizing

C.3 Explainable Financial Reinforcement Learning

We reproduce [16] that compares the performance of DRL algorithms with machine learning (ML) methods on the multi-step prediction in the portfolio allocation task. We use four technical indicators MACD, RSI, CCI, and ADX as features. Random Forest (RF), Decision Tree Regression (DT), Linear Regression (LR), and Support Vector Machine (SVM) are the ML algorithms in comparison. We use data from Dow Jones 30 constituent stocks to construct the environment. We use data from 04/01/2009 to 03/31/2020 as the training set and data from 04/01/2020 to 05/31/2022 for backtesting.

Code available at: https://github.com/AI4Finance-Foundation/FinRL/blob/master/tutorials/2-Advance/FinRL_PortfolioAllocation_Explainable_DRL.ipynb

C.4 Podracer on the Cloud

We reproduce cloud solutions of population-based training, e.g., generational evolution and tournament-based evolution. If GPUs are abundant, users can take advantage of this benchmark to meet the real-time requirement of high-frequency trading tasks. Detailed instructions are provided on our website.

Code available at: https://github.com/AI4Finance-Foundation/FinRL_Podracer

C.5 Ensemble Strategy

The ensemble method combines different agents to obtain an adaptive one, which performs remarkably well in practice. We consider three component algorithms, Proximal Policy Optimization (PPO), Advantage Actor-Critic (A2C), and Deep Deterministic Policy Gradient (DDPG), which have different strengths and weaknesses. Using a rolling window, an ensemble agent automatically selects the best model for each test period. Again on the 30 constituent stocks of the DJIA index, we use data from 04/01/2009 to 06/30/2019 for training, and data from 07/01/2020 to 03/31/2022 for validation and testing through a quarterly rolling window.
D Rights

As the authors of this paper and core developers of FinRL-Meta, we bear all responsibility in case of violation of rights.

Disclaimer: Nothing in this paper and FinRL-Meta repository is financial advice, and NOT a recommendation to trade real money. Please use common sense and always first consult a professional before trading or investing.
Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper’s contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] More computational cost.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] May lead to future works with higher computational cost.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]

2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]

3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [Yes]
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes]
   (d) Did you discuss whether and how consent was obtained from people whose data you’re using/curating? [Yes]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes]

5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]