10-716 Course Project Information

We are providing two datasets for the course project. Students can select which of the two datasets they would like to investigate. This year we have a sponsor, Trexquant, who will award cash prizes to the best teams, and consider students with exceptional projects for internship and full time positions.

Cash prizes:

Tier 1, $5,000

Tier 2, $4,000

Tier 3, $3,000

Senior researchers from Trexquant will be available to answer questions during the project, and during the spotlight conference at the end of the project.

Dataset 1 (Numeric):

Available on box: https://cmu.box.com/s/fl5c538drxsso3ml2lz4gb37guyt572i
(You will need to log in via your Andrew ID)

This is broken into two files, with disjoint sets of 200 alpha signals in each. You are free to combine them, or start with one set, develop your approach, and then test it with the other set.

Use the following to extract the data:

```
> data_array = np.load(file_path, allow_pickle=True)
> data_dict = data_array.item()
```

data_dict keys: 'x_data', 'y_data', 'si', 'di', 'raw_data', 'list_of_data'

x_data: 200 alpha signals, spanning 3 years of data in ~1200 stocks. Each row/data point corresponds to some stock day tuple, and there are 1.123m data points.
x_data shape (1123742, 200)

y_data: target y is next day return.
y_data shape (1123742,)

si: stock index
si shape (1123742,)

di: day index
di shape (1123742,)
start day index: 3776 (corresponding to 20210104); end day index: 4528 (corresponding to 20231229)

raw_data:
11 raw data variables that may be interesting to include in the models and analysis.

list_of_data ['close', 'open', 'low', 'high', 'volume', 'trading_days_til_next_ann', 'trading_days_since_last_ann', 'close_VIX', 'ret1_SPX', 'sector', 'industry']

Dataset 2 (Text and numeric):

https://github.com/Zdong104/FNSPID_Financial_News_Dataset
FNSPID (Financial News and Stock Price Integration Dataset), is a comprehensive financial dataset, containing 29.7 million stock prices and 15.7 million financial news records for 4,775 S&P 500 companies from 1999 to 2023, gathered from four stock market news websites.

The dataset is available at the Hugging Face.

Alternative Misc. Theory Project:

In case any of you do not want to work on the above, and are interested in a pure theory project: we expect this to be a single person team project, and it would not be eligible for the prize above. Please also check with us on the suitability of what you have in mind for your project topic.

Grading

Your project will be worth 24% of your final class grade, and will have 4 deliverables:

1. **Proposal**: 2 pages excluding references (1%)
2. **Midway Report**: 5 pages excluding references (2%)
3. **Presentation**: Spotlight slides presentation (3%)
4. **Final Report**: 9 pages excluding references (18%)

Team Formation

You are responsible for forming project teams of up to two people. Once you have formed your group, please send one email per team to the class instructor list (10716-instructors@cs.cmu.edu) with the names of all team members.

Project Proposal

You must turn in a brief project proposal that provides a precise description of your plans for the course project. Proposals should be at most **two pages long**, and should

include the following information:

- Project title and list of group members.
- Which dataset(s) you intend to analyze
- What methods you intend to use
- A plan of activities, including what you plan to complete by the midway report and how you plan to divide up the work between members.

The grading breakdown for the proposal is as follows:

- 20% for listing datasets and methods
- 80% for plan of activities

The project proposal will be due at **11:59 PM on Monday, March 17**, and should be submitted via Gradescope.

Midway Report

The midway report will serve as a checkpoint at the halfway mark of your project. It should be at most **5 pages long** and should be formatted like a conference paper, with the following sections: introduction, revised plan of activities, preliminary results. The introduction section should be in near final form; the remaining sections will have the results you have obtained, perhaps with placeholders for the results you plan/hope to obtain.

The grading breakdown for the midway report is as follows:

- 10% for introduction
- 30% for a revised plan of activities
- 50% for preliminary experimental results
- 10% for quality of writing

The project midway report will be due at **11:59 PM on Monday, April 7,** and must be submitted via Gradescope.

Final Report

Your final report should be at most **9 pages excluding references**, in accordance with the length requirements for a NeurIPS paper.
The grading breakdown for the final report is as follows:

- 10% for introduction and background
- 70% for results and methods
- 20% for quality of writing (clarity, organization, flow, etc.)

The project final report will be due at **11:59 PM on Friday, April 25,** and must be submitted via Gradescope.

Spotlight Presentation

All project teams will present their work during the last 1-2 classes of the semester. Each team should present a brief (exact time TBD) spotlight presentation, similar to

“spotlight talks” in recent conferences.

The grading breakdown for the spotlight presentation is as follows:

- 25% for topic introduction and motivation
- 50% for survey results
- 25% for presentation clarity and correctness

One of our goals with this was to have the projects creatively differentiate themselves on the methods rather than the data. This would also allow streamlined evaluation and support (and prizes!)

FAQ

Q: Will there be a leaderboard?

A: We don't expect to have a leaderboard. It is more similar to a datathon than a Kaggle challenge; and we will take into account all aspects of the project: performance, novelty of the proposed model, etc.

Q: How will the prizes be determined?

A: The prize will be awarded at the end, and will be largely based on the final report, and spotlight presentation.