Capstone Second Semester Recap

Jean-Gabriel de Montigny

Nicholas Turbide

Members Contributions

	Contributions	%
Jean-Gabriel	- Web Scraper - Sentiment score analysis with LLM - Portfolio Management	50
Nicholas	- Machine Learning development and testing - Hyperparameter testing and automation - Graphical Representation - IBKR Link (WIP)	50

Estimated Time Spent

Task	Estimated Time Spent (Hours)	Contributors
Web Scraping	30	J-G
Sentiment Score (LLM)	45	J-G
Portfolio Management	50	J-G
ML development and testing	100	Nick
Hyperparameters testing	20	Nick
Graphical Representation	10	Nick
IBKR Link (API)	5	Nick
Research – Various Topic - Financial models - Financial equations - Discussions with client	20	60% J-G 40% Nick

Work completed to date:

- Web scraper
    - Scrape financials news from multiple news sites for a specific stock. The scraped content is saved in a CSV file where it can be accessed for further data manipulation.
- Sentiment score analysis

- Using the FinBERT LLM, an LLM based on Google's BERT but trained on financial jargon for higher accuracy, each scraped article is given a sentiment score from negative to positive.

- Portfolio management

  - Using financial equations to optimise how money is invested for a basket of stocks to maximise profits based on historical performance.

  - Used to determine the exit and rebalancing strategy using thresholds. These thresholds can be adjusted for each stock based on historical performance.

- Machine Learning development and testing

  - In depth research and testing to find the machine learning model best suited for our project. We choose to use a LSTM model which is great for the variable market conditions.

- Hyperparameter testing and automation

  - Long testing session to determine the best parameters to use to fit a specific stock based on a multitude of factors.

- Graphical Representation

  - Using the hyperparameters found for a specific stock we use the ML model paired with historical data, financials indicators and the sentiment score collected from the LLM to graph a price prediction for a range of 1-3 days in the future.

- IBKR Link (WIP)

  - The API to connect to the broker we use the execute our trades. Currently works for basic orders but our full strategy is not implemented at this point but will be one of our main priorities as we keep working with out client.

Challenges encountered and resolved:

- Web Scraping
  - Web Scraping was often blocked by "continue reading" or "read more" prompts. After trying multiple methods to web scrape from using specific HTML tags to Selenium to try to click through the articles, we found a site that regrouped news from various news sites and had an HTML structure that made accessing content relatively straightforward. With this method we can scrape news for the past 3 years for most stocks.
- LLM Accuracy

- The first prototype we had for sentiment score analysis used a publicly available LLM with a limited context frame. Because of that limitation we could not put whole article into one context frame which affected the precision of our rating. To solve this issue, we researched for a different LLM to use and found BERT which is an LLM developed by Google's engineers a few years ago. Using BERT improved our results but after a bit more research we found out about FinBERT which is based on BERT but trained with financial data. Switching to that LLM improved the precision of our sentiment score even more.
- ML modelling
    - CUDA architecture (tensorflow issues -> PyTorch)
        - Downloading and setting up CUDA was difficult enough. We also had issues with our tensorflow code setup, since it is a library built for enterprise level code – with emphasis on stability, it uses old versions of CUDA technologies. PyTorch is much more akin to what we needed since it is a research focused library, which can use our GPUs' capabilities.
    - Hyperparameter automation
        - Finding out how to setup a python file to automatically cycle through a list of hyperparameters
        - Follow up: understanding which parameters are even needed to be tested
    - Archetype definition (LSTM vs other types)
        - Scouring through academic papers to find out if there are any benefits to using different models instead of LSTM. Conclusion: for general use LSTM is best suited for our needs
    - **Testing accuracy** (the biggest challenge – essentially endless optimization)
        - Preventing overfitment of the dataset whilst maintaining a reasonable prediction range. This is still an issue; it varies from stock to stock. If it is a predictable stock, like AAPL, it is somewhat fine. But on stocks like NVDA it is somewhat inaccurate due to the rapid movements as of late.

Updates from original proposal:

- Front-end element put on the back burner
    - In the first semester, we worked on a front-end element for our project to make it easier to track market data and easily see breaking news. However, we realised this component was not essential for the success of our project even if it was nice to have. So the front-end development was put on hold until we are satisfied with the fine-tuning of our machine learning model and sentiment score.