Deep Learning for Commodity Price Prediction

Name: Haoming Guo

Student id: 201600384

Supervisor: *Anh Nguyen*

second marker: *Unassigned*

Category:

1. Project Description
2. Aims & Objectives
3. Key Literature & Background Reading
4. Development & Implementation Summary
5. Data Sources
6. Testing & Evaluation
7. Ethical Considerations
8. BCS Project Criteria
9. UI/UX Mockup
10. Project Plan
11. Risks & Contingency Plans
12. References

Project Description

This project will develop and implement a machine learning program in python which is able to predict the commodities price (especially the gold) based on the history time-stamp stock data and other helpful indicator. The program will return a real number which represent the future price rather than a prediction for only a direction. I will develop and implement a program after learning some relevant papers and try to combine the advantages from different models to make the prediction having the highest precise. The finished project will predict the commodities price successfully and precisely when it is given the enough history data. I also hope the project will still have the ability to predict the relevant commodity not only for the gold.

Aims & Objectives

The main aim of my project is to develop a python program to implement a relatively professional model to predict the stock price like gold. The output of the model is a number which indicate the future price rather than a relative change direction. It is much more powerful than a simple classification model, which can not only show you the direction but also help people to decide the amount when they go short or long. This project will allow everyone to use and run it and get a relatively precise price which can be used as one of the indicators to help people make a right financial decision.

This project has several objectives to judgment and make sure it can be a successful program and model. I believe the first and the most important objective for prediction is precision. Any prediction without an enough high precision is useless. Another objective is that the model can read and use other indicators which is helpful and informative to enhance the performance rather than only extract the features from the time-stamp stock price which may include the intraday return, return with respect to close and open price, Dow Jones Industrial Average and Moving Average. The third objective is speed. The project should have a reasonable speed to allow users run multiple times with a low time cost. Timeliness and low cost are also a successful model need have in the financial market.

Key Literature & Background Reading

There are some paper and materials did actually help and inspire me a lot. The first one named *Forecasting directional movements of stock prices for intraday trading using LSTM and random forests* [2]. They use random forest and LSTM respectively as training methods to predict the intraday trading price S&P 500 from January 1993 to December 2018. The idea in the feature selection step gives me a strong support. First of all, they have different feature generation for different model. For the random forest, there are three signals: intraday returns, return with respect to last closing price, return with respect to opening price. Apply these three signals to the specific day to produce 93 features. As for LSTM, they input the model with 240 timesteps and 3 features applied with Robust Scaler standardization to make it robust to outliers to predict the direction of No.241 intraday return. It inspires me to not only use multiple features or indicators as inputs, but also need to apply some specific methods like standardization to make data robust. The detail about the training specification like loss function, optimizer, batch size and number of decision

trees all have reference significance. As the result, the performance of using “intraday” is better than using "NextDay” for setting. And the LSTM outperforms than the random forests. This information definitely will be helpful to my work since my aim is to develop a model which can outperform than others.

Development & Implementation Summary

The development environment is not sure yet. However, since the implementation language I chose is python, the final environment will be one of the choices listed below. First candidate is the *PyCharm,* which is a famous Python IDE. And another choice is using *Colab*, since maybe the running such a model is too heavy for my own PC or MacBook and *Colab* allow me to run the model much faster by using a more powerful GPU they distribute to me. As for the reason for why I chose Python as the implementation language, there have a lot of reasons I can list. But one of the most important points is that python have many powerful libraries and implemented functions which I can call and use conveniently. Another reason is that many similar projects are implemented by Python, much more experience and examples can support me finish this project successfully. Because it is a totally new project and topic for me, there is no doubt that I will meet lots of problem, no matter what kind of problem, both for code or knowledge.

As for the key questions about how to implement the model, I believe I can solve it quickly based on my previous experience about such a deep learning project. The process will follow such an order: define the hyperparameters, implement the model structure, train and test it finally. and the structures about the neural network, I plan to firstly refer the specific structure from the papers or websites. Then do some slight change and test it to try my best to achieve a better performance.

Data Sources

The dataset is from the website, Kaggle, which is famous website for data science and its project and competitions. The link is https://www.kaggle.com/datasets/prasertk/historical-commodity-prices-from-20002022. This dataset has the prices from 2000-2022 of various commodities, including gold, palladium, nickel, brent oil, nature gas and wheat. It is a publicly available dataset based on its License: CC BY-SA 4.0, which allow me to share and adapt the dataset. It will not involve or violate any confidentiality of personal information since the topic and content of this project and dataset is all about information of commodities’ price which is belong to public information and everyone can get it.

In another paper [3], the author defines the prediction problem for gold as a classification problem. Although it is not fit with my objective which predict the absolute price tomorrow, it is still many thinking and methods which can support my work. One important point is that the author tries to figure out whether the intermarket financial variables like stock index price, exchange rates, bond rates and other commodity price can affect forecasting the price of gold.

Testing & Evaluation

Since this project will predict the commodity absolute price, test and evaluation is a necessary step in the implementation. Firstly, I will divide the dataset to three subsets: used for feature generation, training and testing separately. The testing dataset is used to test the precision of the predict result. Furthermore, I will test and evaluate every parameter combination like batch size, learning rate, and feature selections in order to get a best result, which can also be called as white box test. After determine the parameters, I also plan to execute some black box test like using different but relevant commodity price as input to evaluate the performance of the model.

Ethical Considerations

I state that I have read our ethical guidelines and will follow them. This project is a personal project and will not involve any other human participants exclude me. Before begin this project, I have obtained formal ethical approval from supervisor. The Central University Committee have the right to suspend or discontinue my work anytime when it is conflict with the conditions of its original approval. All the data used is genuine and I have been allowed to use it. All the data analysis will be completed by myself. The project will be implemented and finished in the UK and involve nobody else. I won't pay others to take part in my project.

BCS Project Criteria

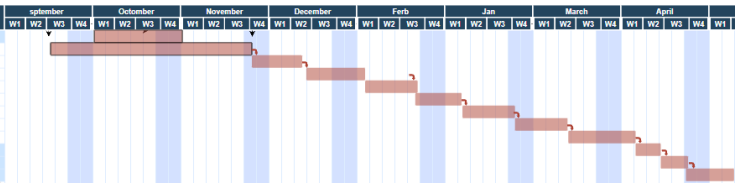
My project needs me to using python and its powerful libraries to implement a deep learning model which can read and analysis the commodity price and predict the future price finally, which can meet the first criteria: apply practical and analytical skills gained during the degree program. As for the innovation and/or creativity, I believe it can also can be found during my work since my final aim is trying to find a model which can outperform than other previous models. I need analysis the advantage and disadvantage of previous models and create my own model. I will synthesis information, idea and practices from internet and other people to solve my task with a detailed evaluation successfully. Furthermore, my project has a wider context since the market activities are happening anytime, anywhere. And my project allows people use the powerful calculate ability of the computer to solve one of the hardiest tasks in the financial market, the price prediction. Besides, there have a detailed plan to manage the tasks and push myself to complete the work. Any thinking and ideas about the project and self-evaluation will be write down as the Dr Stuart suggested in the lecture of week one.

UI/UX Mockup

Since this project will be implement in the python environment and maybe will not be package as a website or an application in the PC or phone, so there will not have the question about the design and implement about the user interface (UI). As for the user experience, I plan to encapsulate most of the code in the proper method which is easy for user to read and call these functions. According to the objective, most of user can run and get their result with the help of the instructions and comments.

Project Plan

	Task Name	Duration	Start	End
1	finish the Detailed Proposal	35 days	2022/10/01	2022/10/30
2	Background Reading	1 day	2022/09/15	2022/11/20
3	implement and evaluate the LSTM	1 day	2022/10/21	2022/11/14
4	implement and evaluate the random forests	1 day	2022/11/15	2023/01/03
5	implement and evaluate the Logistic Regression	2 days	2023/01/04	2023/01/21
6	implement and evaluate the a SVC model from scikit-learn	1 day	2023/01/22	2023/02/08
7	design my own model	3 days	2023/02/09	2023/02/25
8	develop my own model	27 days	2023/02/26	2023/03/15
9	optimizition work	1 day	2023/03/16	2023/04/06
10	test and evaluate the model	20 days	2023/04/07	2023/04/15
11	Make project Video	20 days	2023/04/16	2023/04/21
12	finish the project Dissertation	20 days	2023/04/22	2023/05/12



Risks & Contingency Plans

Risks	Contingency plan	Likelihood	Impact
Hardware failure	Using several devices synchronously. Use *Colab* when the hardware cannot support and finish the task.	medium	low
Software failure	Save and update the copies in the different devices.	medium	high
Running out of time	Use more leisure time and vacation to catch the plan	low	high
Programming problems	Seek the solution and help from the internet (Stack overflow) or from the supervisor and friends	high	low
Software or environment version doesn't match and cause some code problems.	Unify environment version number and post it in the beginning. Double check this kind of information before running the code.	medium	medium

References

[1]: P. Kanawattanachai, "Major commodity prices from 2000-2022," Kaggle, 09-Apr-2022. [Online]. Available: https://www.kaggle.com/datasets/prasertk/historical-commodity-prices-from-20002022. [Accessed: 30-Oct-2022]

[2]: P. Ghosh, A. Neufeld, and J. K. Sahoo, "Forecasting directional movements of stock prices

for intraday trading using LSTM and random forests," arXiv.org, 30-Jun-2021. [Online]. Available: https://arxiv.org/abs/2004.10178. [Accessed: 30-Oct-2022]

[3]: M. Potoski, "Predicting gold prices - stanford university." [Online]. Available: http://cs229.stanford.edu/proj2013/Potoski-PredictingGoldPrices.pdf. [Accessed: 30-Oct-2022]