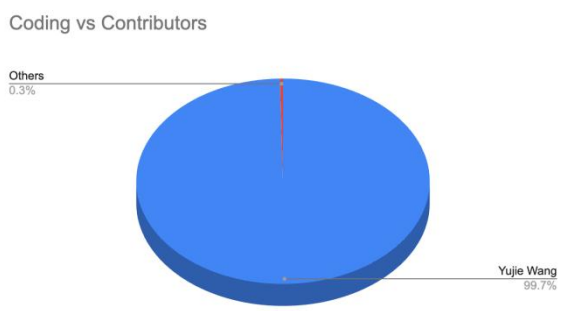
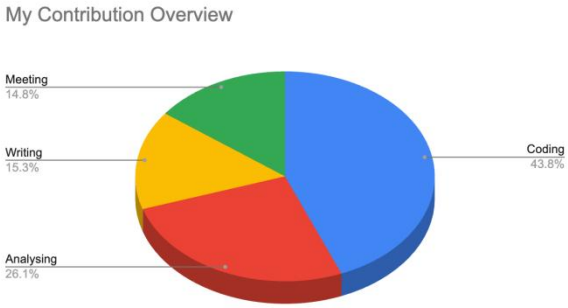
Data Science Mini-Project 2022/23 - Individual Reflective Report

Yujie Wang – 24 May 2023

Summary

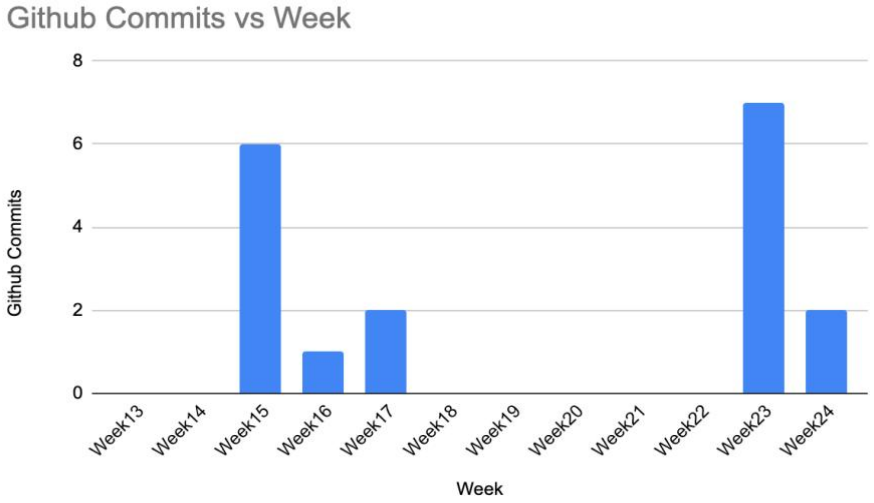
I contributed to approximately 50% of the overall project effort. The contributions were primarily from two members. My contributions were spread across coding, analysis, writing, and meeting. I attended all the group sessions. I contributed to most of the coding activities as shown on Github. For the written report, I involved in most parts of Methodology, Data Preparation, and Results.(All contents concerning with the ARIMA Model, LSTM Model and Mid-Price predictions).And the parts about Mid-Price tasks were all done by myself. Furthermore, I gave 1/3 of the oral formative and summative presentations. I consistently contributed every week via either Github or WhatsApp (and sometimes via Teams). In addition to this information, I have attached my weekly activity logs.



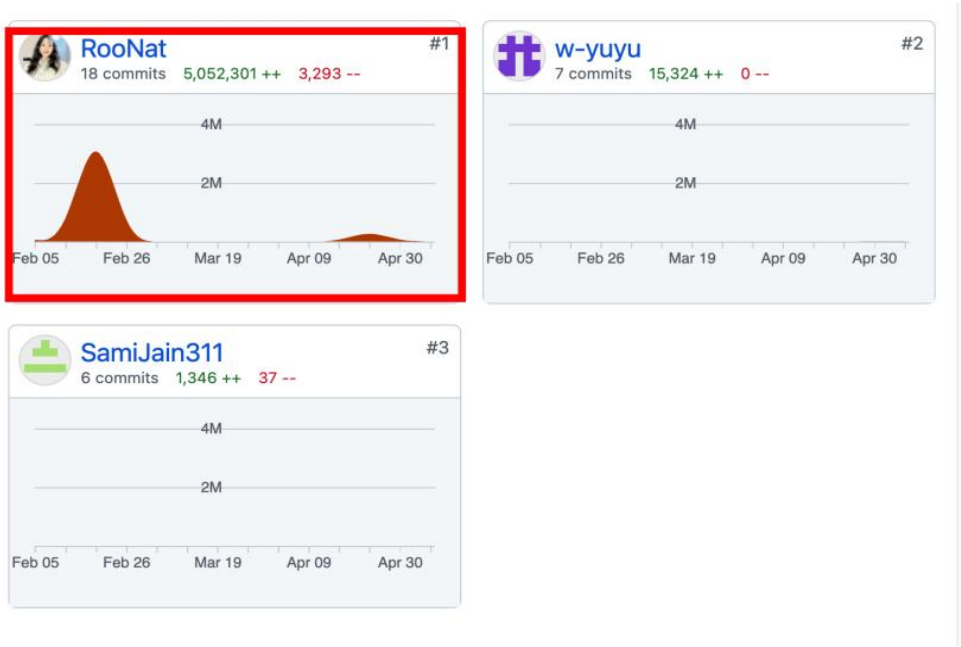


Github Details

I did have some activity on Github, where I primarily added data and codes. Many of the images were derived from Excel.
Our coding work includes the data processing, EDA, predicting stock price with ARIMA model and LSTM model, predicting the Mid-Price with ARIMA model and other models, predicting the Mid-Price changes with various models (including two-class classification problemand thed three-class problem), and the simulators. I did almost all the coding work except EDA, simulators, and using the LSTM model to predict the stock price.

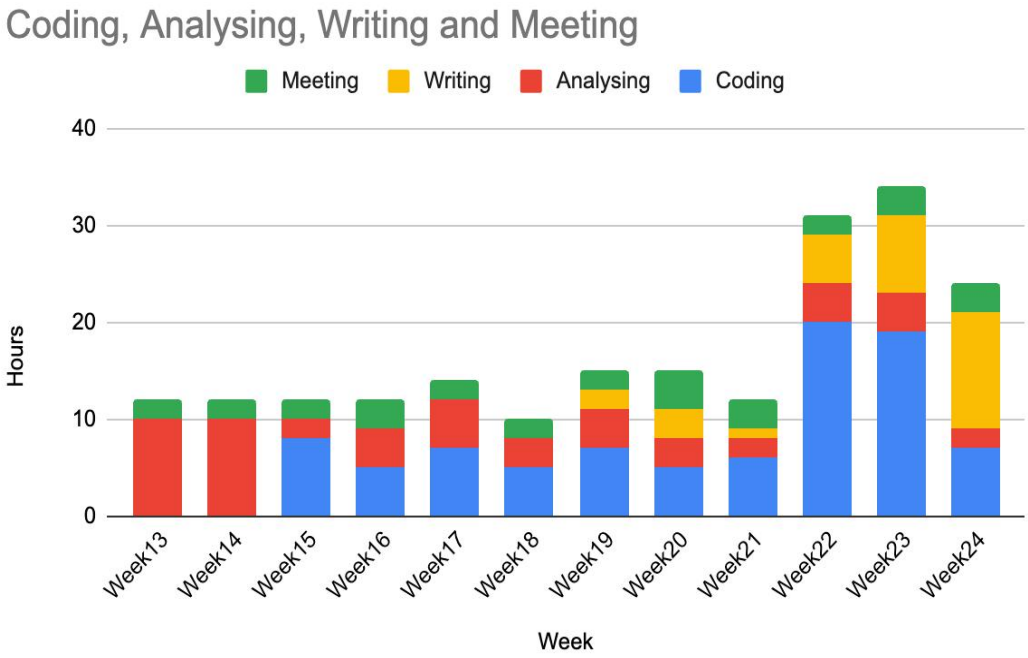


Based on the records shown on the Github, I contributed to 18 commits, including 5052301 additions and 3293 deletions.



Activities

For the first several weeks, I spent most of the time on reading paper, data processing and analysing the data format. Basically, we had the meetings once or twice a week. I always performed actively over the period. After the spring break, I spent most time on feature engineering and model optimizing. Therefore, I did provide abundant papers and materials for members to write the report.



Appendices： Log Activities

DSMP progress report

Student Name:Yujie Wang

UoB Username: ni22398@bristol.ac.uk

Week	Topic	Notes
15	Individual tasks:	Process the data
	Collaborative tasks:	Process the data,Find the materials about LOB,and discuss about the algorithm
	Challenges:	How to caculate the profit How to find a good algorithm to predict
16	Individual tasks:	has related the .txt file with .csv file and process the data
	Collaborative tasks:	Find the model we're going to use and discuss about the next step we're going to do.
	Challenges:	Choose the correct model to predict.
17	Individual tasks:	I have processed all of the given data currently,. And I'm working on learning ARIMA time series algorithm to train the data. At the same time, I'm trying to visualize the data in limt order book.
	Collaborative tasks:	Determine several algorithms we're going to working on, including ARIMA, LSTM, Logistic regression and so on.
	Challenges:	I'm still not familiar with the relevant finantial concepts, so it's hard for me to leverage the data correctly to train in the model. And I'm not sure about how the stock market work.
18	Individual tasks:	Read the relevant materials about ARIMA model
	Collaborative tasks:	We separately learned the different models about dealing with the LOBs data and OHLC data to predict
	Challenges:	The relevant professional knowledges about ARIMA

		model, time series concepts
19	Individual tasks:	Using ARIMA model to train the OHLC data and predict primarily
	Collaborative tasks:	We discuss about how to leverage the LOBs data to model and predict, keep learning the algorithms
	Challenges:	How to leverage the LOBs to make classifications and real-time predictions, I'm going to apply the LOBs data to the ARIMA model.
20	Individual tasks:	Make the sildes and build the ARMA model and ARIMA model to simulate and make initial predictions.
	Collaborative tasks:	Prepare the ppt and do the presentation
	Challenges:	Again, take the input LOBs data into ARIMA model, It's better to use ARIMA than to use more difficult models like LSTM. So we need to optimize the ARIMA model.
21	Individual tasks:	calculate the mid-price of the limit order books data
	Collaborative tasks:	Formative written report and do the mid-price predictions
	Challenges:	Optimize the ARIMA model and using more features to do the predictions.
22	Individual tasks:	Use ARIMA Model to predict the Mid-price, and improve the model for OHLC.
	Collaborative tasks:	The EDA, ARIMA model for Mid-Price, and LSTM for OHLC
	Challenges:	How to normalize the data when using LSTM model for mid-price prediction. Do we need to replace 0 with NAN values, since the levels are different at the dfiferent timestamp
23	Individual tasks:	Mid-price movement classification with CNN and LSTM, Mid-price predictions with LSTM, Mid-Price simple classfier with RNN. Mid-Price predictions with

		ARIMA model.Write the code about transforming all of the LOBs data to the .csv format and calculating relevant benchmark.
	Collaborative tasks:	My individual tasks + prepare for the simulater+ processed all of the data with AWS+ EDA+ OHLC predictions with LSTM
	Challenges:	Also the questions about the format of features and optimizing the models.
24	Individual tasks:	I have done all of the code work and I'm checking and modifying them, include predictions with 1. Data Preprocessing 2. OHLC data and ARIMA Model, the returns predictions with OHLC data ARIMA Model 3. The mid price prediction with ARIMA model 4. The mid price prediction with LSTM model 5. The mid price classification with RNN model 6. The mid price multiple-class classification with CNN and LSTM model. And then I'm writing my report about mid price prediction.
	Collaborative tasks:	Write the summative report and do the simulator
	Challenges:	There are too much contents and I need to organize them and to see which part I need to put them on the report. Other than that, I think my model doesn't work very well, I feel like there's some space to make progress.