Stock Price Prediction and Optimal Portfolio Construction using Quantitative Research — Progress Report 2

Tarusha Silva, *Sylvia* Zhang, *Sergey* Lebedev, and *Thomas* Michel

[1] *Columbia University in the City of New York, MSOR, Fall 2022*

Abstract. The main goal of this project is to use financial and alternative data in order to participate in an internal real-time forecasting competition. In this report, we highlight several of our approaches in predicting prices for a portfolio consisting of 50 stocks, 50 ETFs and 10 crypto currencies to construct the optimal portfolio among our peers. On this paper we will focus on explaining the main data-science workflow of generating ideas, information sourcing, features extraction, tuning models, strategies simulation, and performance's evaluations. Starting from a viewpoint of assuming asset returns are random and normally distributed, we will also discuss approaches that utilize Geometric Brownian Motions, LSTM, other the machine learning methods, optimization, data scraping, decision analysis, sentiments analysis used in our approach and will try to analyze our results. To date, our portfolios have performed quiet well compared to peers and in this report we discuss our findings and focus areas for future research and finetuning of the model.

1 Introduction

Advanced machine learning techniques are increasingly replacing and refining the fundamental approach to asset pricing and portfolio construction in the marketplace. The modern investor must be fluent in these approaches to exploit market inefficiencies and generate alpha in their portfolios. In this report we highlight our approach to generate market views and forecasted asset prices to build the portfolio that maximizes risk adjusted return in order to win a 14-week long competition among our peers.

There are many approaches to this problem in the market, but most of them are not known to the average investor given the proprietary nature of the findings. However, our team is well-versed in techniques and methodologies spanning the fields of Finance, Operations Research and Machine Learning. We endeavor to use this knowledge in formulating the ultimate algorithm that will utilize a variety of external data sources and advanced Machine Learning techniques to give us a layout for the best-performing portfolio among our peers.

The best algorithm will use the least amount of external data sources to generate future asset prices to improve the efficiency in the runtime and to reduce potential biases in the construction of the algorithm. The best algorithm will also induce the least amount of strain on the computing infrastructure given that our main limitations are in terms of hardware to run our code.

Considering all of the above, our goal with this project is to formulate an algorithm that exploits market inefficiencies and accurately predicts the prices of 150 selected financial assets week-after-week while circumventing our limitations in data generation and computing hardware.

2 Strategies and Methods

2.1 Week 1

For the construction of our portfolio in the first week, we used a simple approach where we assumed that each asset, regardless of which class they belong to, its returns are normally distributed with a mean value of zero and a standard deviation corresponding to its historical price movements.

Given that the portfolio consists of 10 crypto currencies, some of which originated recently, we used a historical window length of 1 year spanning September 09, 2021 to September 09, 2022.

We generated historical time series of the asset prices using Yahoo Finance. After obtaining this dataset, we calculated the weekly returns for each asset. This modified dataset was then used to calculate the standard deviation of returns for each asset. Having gathered all this information, we then ran 1,000 simulations for each asset to chart various paths the asset might take in the upcoming week. We then took the average of these simulations as our forecast for each asset. These forecasts were then appended to the historical price dataset to be input as a time series in the PyPortfolioOpt library to generate the optimal portfolio that maximizes the Sharpe ratio.

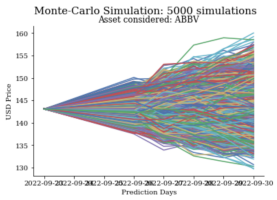

Figure 1. ABBV

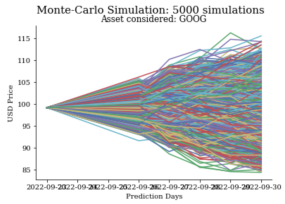

Figure 2. GOOG

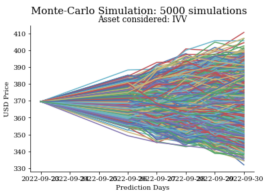

Figure 3. IVV

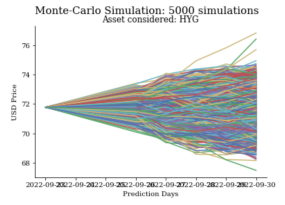

Figure 4. HYG

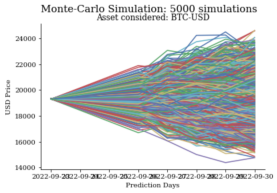

Figure 5. BTC-USD

Furthermore, we developed a ranking system where for each simulation epoch we collect the return generated by each asset and then rank them into quintiles. The number of times each asset falls into these quintiles is then used to calculate the percentage of the asset's returns that fall into rank 1,rank 2,rank 3,rank 4, and rank 5, where rank 1 and rank 5 are representative of worst performance and best performance, respectively.

2.2 Week 2

In the second week, we took a slightly different approach and used the theoretical framework that asset prices follow a Geometric Brownian Motion to forecast future prices. Also in contrast to the first week, we took daily price returns to generate probable paths that each asset might take in the upcoming week. We simulated prices for each asset this way 5,000 times and took the mean of those simulations as the forecasted price for the asset.

These new prices were then appended to the historical price dataset as before to be used in optimizing the portfolio using PyPortfolioOpt with the Sharpe ratio as the decision criteria.The ranking system for each class was the same as the first week.

2.3 Week 3

We tried two time series models in the third week, the Autoregressive Integrated Moving Average model (ARIMA) and the Autoregressive Moving Average model (ARMA). We trained the two models with data split before 06/16/2022 and validated them with data after 06/16/2022. We then compared two models' performance and adopted the one with smaller average Root Mean Square Error when forecasting future prices. The model selected turned out to be ARMA.

Similar to what we did in previous weeks, we appended the predicted prices to the historical prices and then optimized the portfolio using PyPortfolioOpt with the Sharpe ratio as the decision criteria.

2.4 Week 4

For the fourth week, we took a different approach. First, we created 3 indexes to represent the 3 asset classes in our investment universe, i.e. an index for stocks, ETF and crypto currencies. This was done to help us identify external datasets that may have predictive power by calculating their correlation with the individual asset indexes.

The following graphs chart the indexes of the 3 asset classes over time:

Figure 6. Stocks

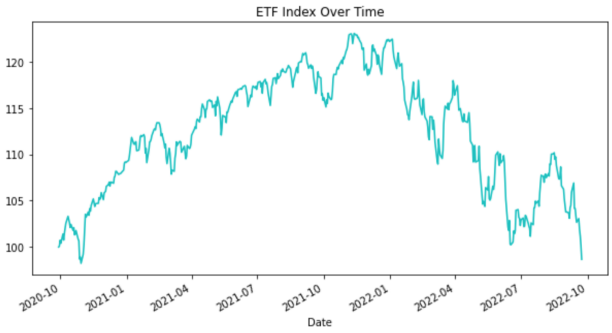

Figure 7. ETFs

To begin with, we evaluated the correlation of these 3 indexes with the VIX index, a dataset of commodity prices spanning 23 commodities, and the individual indexes themselves.

The results for the commodities and indexes are as follows (truncated version to show datasets with highest

Figure 8. Cryptocurrencies

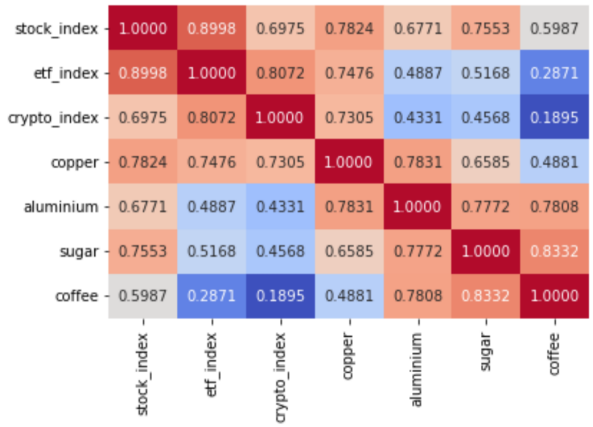

Figure 9. Correlation matrix for commodities

correlation):

The correlation matrix for the VIX index and the 3 asset indexes are as follows:

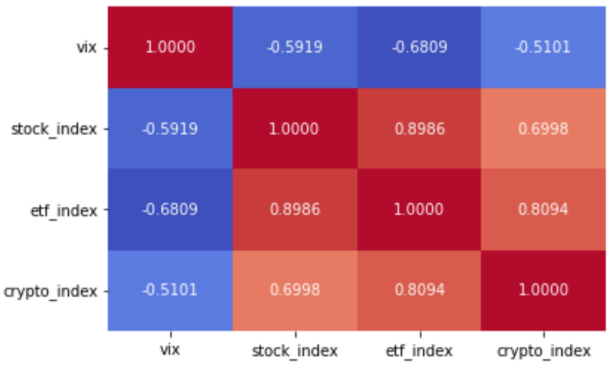

Figure 10. Correlation matrix for VIX, stocks, ETFs and cryptocurrencies

We used this correlation matrix to identify the external data sources that may be useful in predicting individual assets within the 3 classes. For example, we concluded that for the individual stocks, the prices of coffee, copper and sugar seemed to be useful along with the value of the VIX index.

After identifying these datasets, we used a LSTM model to forecast the prices for each asset in the upcoming week. These values were again appended to the historical price dataset to generate the optimal portfolio using PyPortfolioOpt.

The ranking system for each asset was modified now, however, this week's submission was also different due to the fact that we shorted some assets as well. In all previous portfolios we held long positions.

2.5 Week 5

We did not change our model here given how well it performed in the last week's competition. Therefore, we just retrained the model with more up to date values of the same external datasets we used last week to generate the weights for each asset in the optimal portfolio.

2.6 Week 6

During this week, we faced an issue where the provider of our dataset of commodity prices spanning 23 commodities had not updated their prices for the most recent week. We decided to circumvent this issue by using other external datasets that we believe reflect the movements in global commodity prices. Therefore, in addition to the values for the VIX Index, we decided to integrate the values for SP 500, Effective Federal Funds Rate, and 5 Year Forward Inflation Rate in our LSTM model to predict the prices of assets in our investment universe and determine their respective weights in the optimal portfolio.

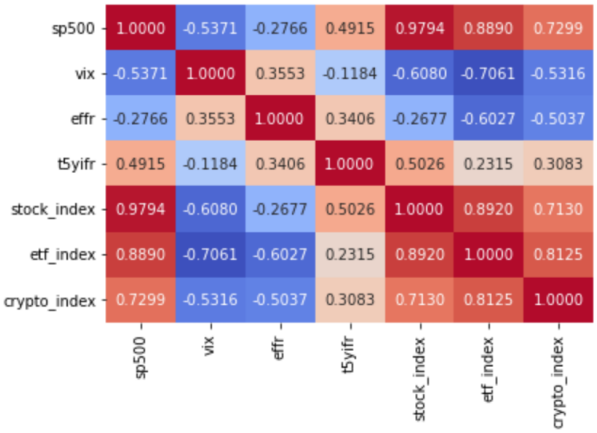

Figure 11. Correlation matrix of the 3 asset indexes with the new external datasets

2.7 Week 7

We were still facing the issue with our dataset of commodity prices spanning 23 commodities not being updated to reflect their most recent prices. Our approach to navigating this still remained the same with the use of the VIX Index, Effective Federal Funds Rate, and 5 Year

Forward Inflation Rate in our LSTM model.

However, we augmented the allocation of weights in our portfolio further this week by identifying stocks that had earnings releasing during our forecast period. If we identified any stock that had such an event coming up (using web scraping methodologies), we would increase its weight by allocating a percentage of the weight of another stock that did not have any event coming up during the same period. To be consistent with how this exchange of weights happens among the stocks, we first separated the stocks in the portofolio to ones with long positions and ones with short positions. Afterwards, we identify the first stock with an event coming up and the other stock which has a similar long/short position and no event coming up but with the highest absolute weight among those second type of stocks. Next, we allocate 5% of the weight from this second type of stock to the stock we identified to have an event coming up during the period and repeat the process.

2.8 Week 8

Learning from our mistakes during the past week, where we missed the correlation of the cryptocurrencies with mainstream currencies, we augmented our external datasets with two new datasets: The US Dollar Index and the GBP/USD exchange rate. Afterwards, we ran the same

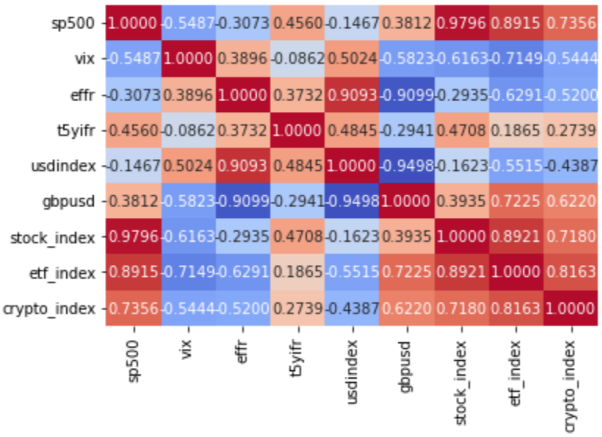

Figure 12. Correlation matrix of the 3 asset indexes with these new external datasets

algorithm that was used in the previous week to determine the weights for our optimal portfolio.

2.9 Week 9

The same algorithm as the previous week was used given the exceptional performance. However, we have exhausted our event-driven reallocation of weights among the 50 public equities given that the earnings season has ended.

We, however, modified our algorithm for ranking by utilizing a simulation-based approach where we once again assumed asset prices behave as Geometric Brownian Motions. We ran 5,000 simulations and for each simulation epoch we collected the return generated by each asset and then ranked them all into quintiles. The number of times each asset falls into these quintiles is then divided by the number of simulations to calculate the percentage of the asset's returns that fall into rank 1, rank 2, rank 3, rank 4, and rank 5, with rank 1 and rank 5 representing the worst performance and best performance, respectively, in each epoch.

2.10 Week 10

This week, we were able to regain access to the external dataset spanning 23 commodities that was updated for the end of the previous week. However, rather than dispensing the modified external dataset we have created in the absence of it, we chose to append this commodities price series into it and run our base LSTM algorithm as described above.

The ranking algorithm remained the same as last week.

2.11 Week 11

Given our confidence in the model, we used our base algorithm one more time, for both portfolio construction and ranking, to generate the optimal portfolio weights and rankings.

3 Results

3.1 Week 1

The individual assets that formed the optimal portfolio in Week 1 along with their respective weights are as follows:

Our ranking among peers and performance in terms of

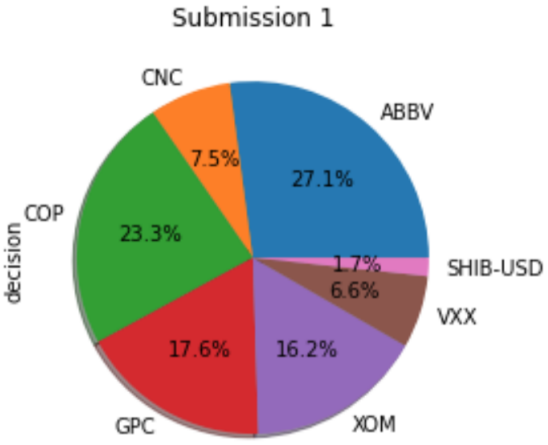

Figure 13. Weight amongst asset for week 1

forecasts and portfolio return as per the stipulated ranking system are as follows: **Rank 1**.

3.2 Week 2

The individual assets that formed the optimal portfolio in Week 1 along with their respective weights are as follows:
Our ranking among peers and performance in terms of

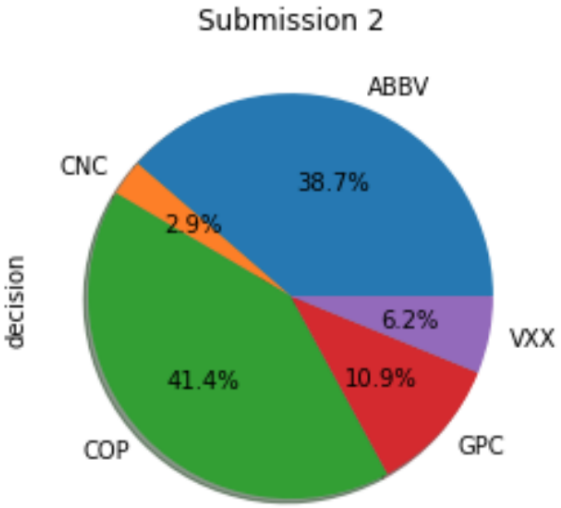

Figure 14. Weight amongst asset for week 2

forecasts and portfolio return as per the stipulated ranking system are as follows: **Last Place due to error in submission file**.

3.3 Week 3

The individual assets that formed the optimal portfolio in Week 1 along with their respective weights are as follows:
Our ranking among peers and performance in terms of

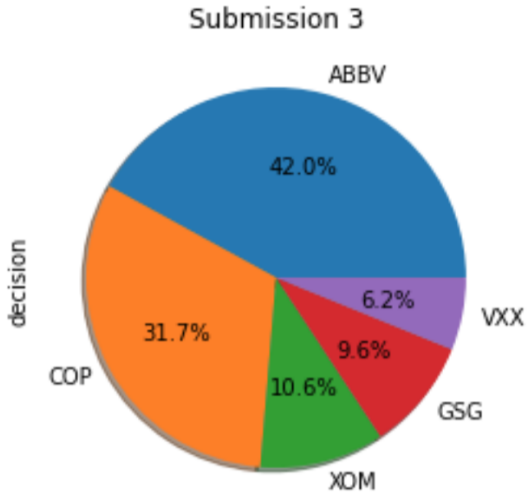

Figure 15. Weight amongst asset for week 3

forecasts and portfolio return as per the stipulated ranking system are as follows: **Rank 2**.

3.4 Week 4

The top 20 assets with the largest absolute weights of the optimal portfolio in Week 4 are as follows:
The only asset with zero absolute weight in this portfolio

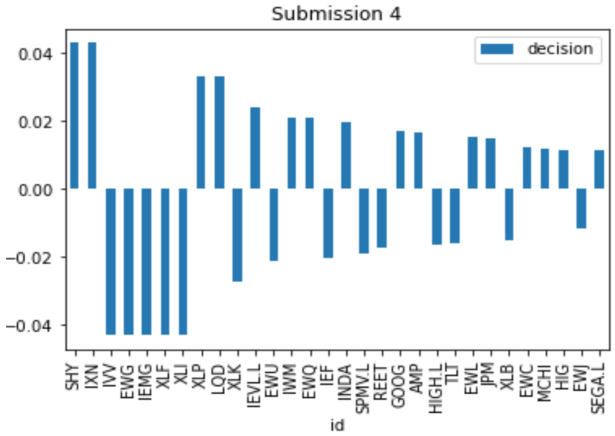

Figure 16. Weight amongst asset for week 4

was SHIB-USD.

Our ranking among peers and performance in terms of forecasts and portfolio return as per the stipulated ranking system are as follows: **To be determined as of October 5th**.

3.5 Week 5

The weights are largely the same as the previous week given that the algorithm used only one more week of new data and no significant events occurred in the markets over this period.

Our ranking among peers and performance in terms of forecasts and portfolio return as per the stipulated ranking system are as follows: **Rank 1**.

3.6 Week 6

The top 20 assets with the largest absolute weights of the optimal portfolio in Week 6 are as follows:
Our ranking among peers and performance in terms of forecasts and portfolio return as per the stipulated ranking system are as follows: **Rank 1**

3.7 Week 7

The following bar chart shows the portfolio weights of the 50 stocks before the exchange of weights happens to account for events happening during the forecast period. The bars in red identify stocks with an event coming up during the forecast period:
The next bar chart shows the portfolio weights of the 50 stocks after the exchange of weights happen to account for events happening during the forecast period.The bars in red identify stocks with an event coming up during the

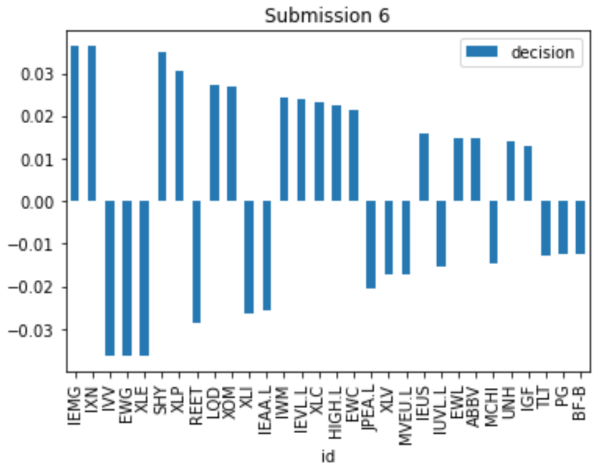

Figure 17. Weight amongst asset for week 6

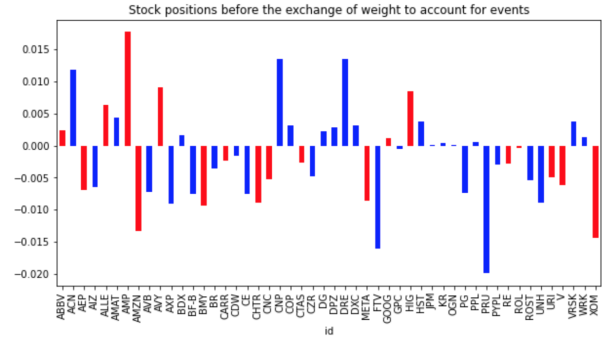

Figure 18. Weight amongst asset for week 7

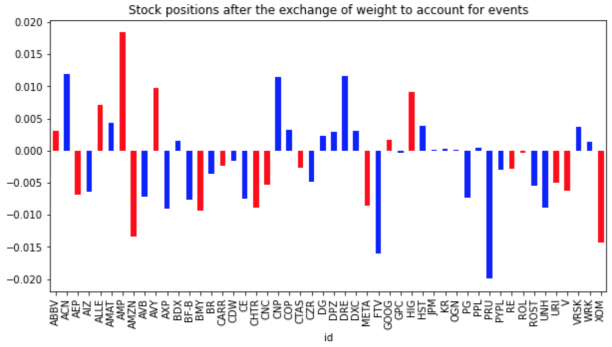

Figure 19. Weight amongst asset for week 7

forecast period:

Our ranking among peers and performance in terms of forecasts and portfolio return as per the stipulated ranking system are as follows: **Rank 4**.

We were penalized by our short positions in the10 cryptocurrencies of our investment universe as they rose over this period given the turbulence surrounding the U.K. Pound.

3.8 Week 8

The following bar chart shows the portfolio weights of the 50 stocks before the exchange of weights happens to account for events happening during the forecast period. The bars in red identify stocks with an event coming up during the forecast period. The algorithm seem to have placed disproportionately higher weights on stocks in the Energy sector:

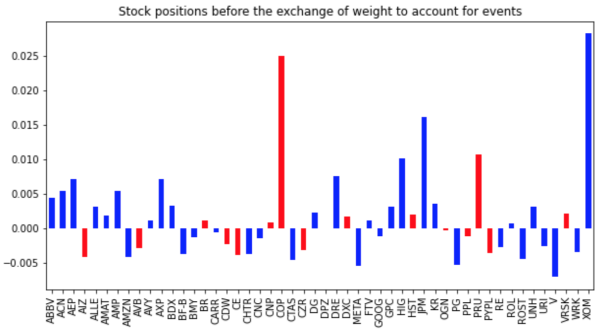

Figure 20. Weight amongst asset for week 8

This following bar chart shows the portfolio weights of the 50 stocks after the exchange of weights happen to account for events happening during the forecast period.The bars in red identify stocks with an event coming up during the forecast period:

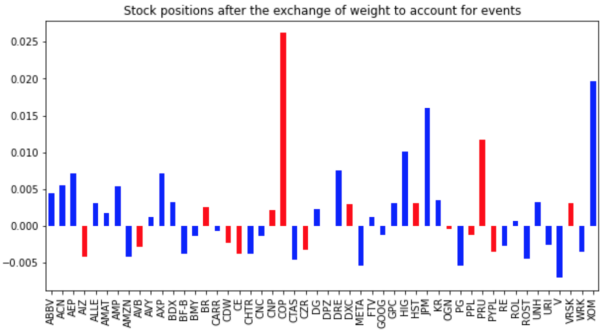

Figure 21. Weight amongst asset for week 8

Our ranking among peers and performance in terms of forecasts and portfolio return as per the stipulated ranking system are as follows: **Rank 1**.

3.9 Week 9

The following bar chart shows the portfolio weights of the 50 stocks before the exchange of weights happens to account for events happening during the forecast period. The bars in red identify stocks with an event coming up during the forecast period. Only 3 stocks have events coming up in the forecast period, indicating that we have likely exhausted our event-driven reallocation of weights

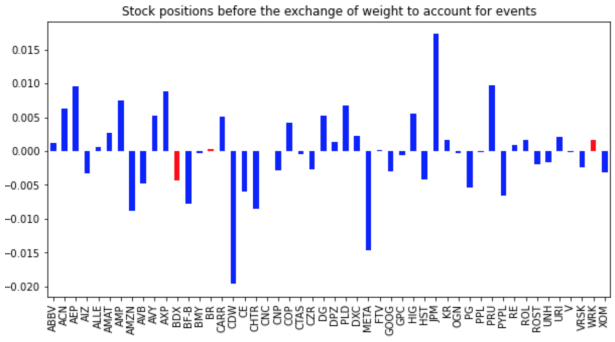

Figure 22. Weights of the 50 public equities in the portfolio separated into those with events (red) and no events (blue) during the forecast period before the reallocation of weights for week 9

among publicly traded stocks.

This following bar chart shows the portfolio weights of the 50 stocks after the exchange of weights happen to account for events happening during the forecast period.The bars in red identify stocks with an event coming up during the forecast period:

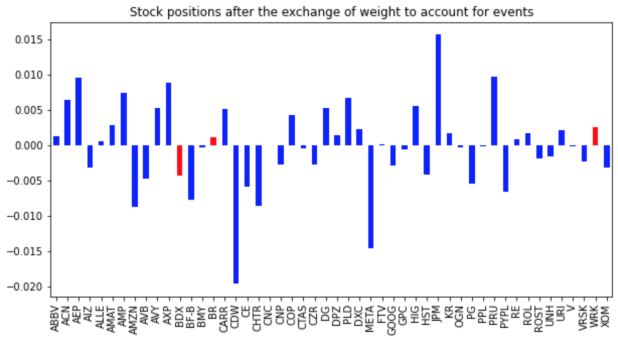

Figure 23. Weights of the 50 public equities in the portfolio separated into those with events (red) and no events (blue) during the forecast period after the reallocation of weights for week 9

This week, we were heavily affected by company-specific risks where there were mass layoffs in the technology sector and volatility in the cryptocurrency market after the bankruptcy of the FTX trading platform. These resulted in some of our assets behaving in the opposite direction from what our algorithm predicted, impacting our overall ranking among peers.

Our overall ranking among peers and performance in terms of forecasts and portfolio return as per the stipulated ranking system are as follows: **Rank 9**.

3.10 Week 10

The attribution of the portfolio's returns among the individual assets, separated by asset class, are shown in the following charts:

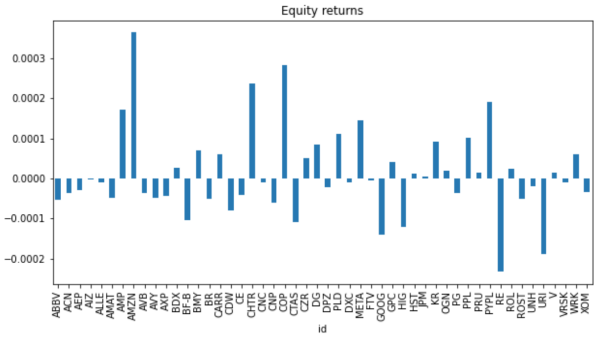

Figure 24. Attribution of Week 10 portfolio return among public equities

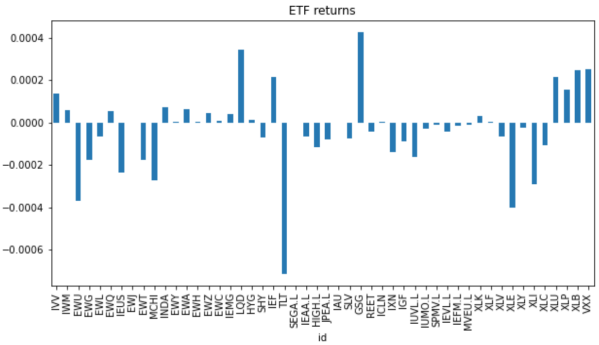

Figure 25. Attribution of Week 10 portfolio return among ETFs

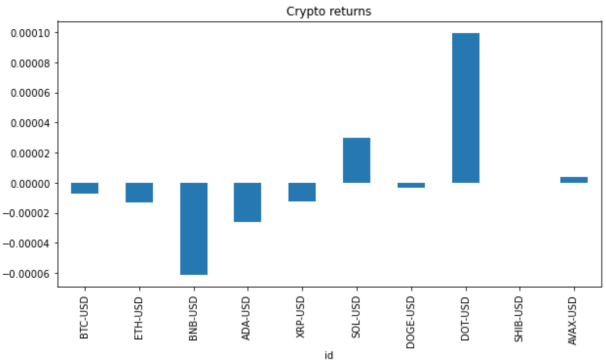

Figure 26. Attribution of Week 10 portfolio return among cryptocurrencies

Our overall ranking among peers and performance in terms of forecasts and portfolio return as per the stipulated ranking system are as follows: **Rank 7**.

3.11 Week 11

The attribution of the portfolio's returns among the individual assets, separated by asset class, are shown in the following charts:

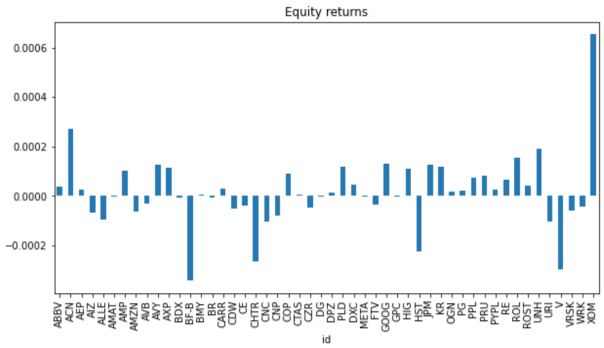

Figure 27. Attribution of Week 11 portfolio return among public equities

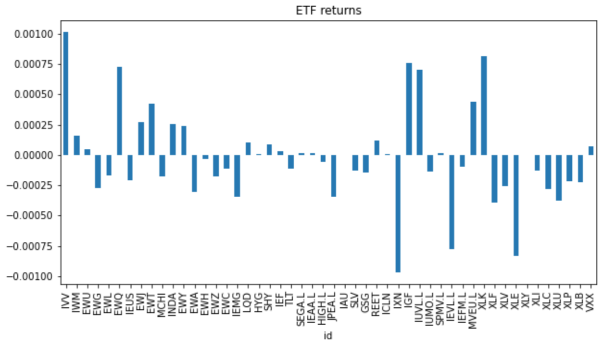

Figure 28. Attribution of Week 11 portfolio return among ETFs

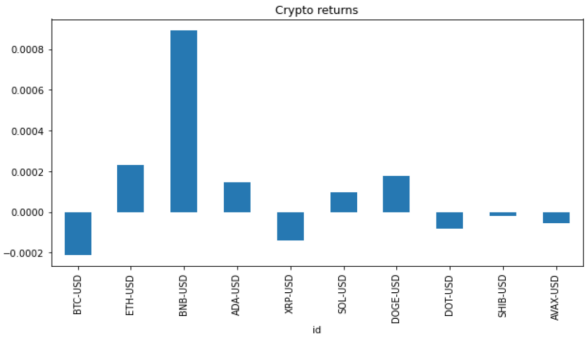

Figure 29. Attribution of Week 11 portfolio return among cryptocurrencies

Our overall ranking among peers and performance in terms of forecasts and portfolio return as per the stipulated ranking system are as follows: **Rank 1**.

4 Discussion

Starting from a purely theoretical framework, our models and algorithms have developed to take on more practical approaches in predicting asset prices. In contrast to many of our peers, we have let the model or machine make a lot of our decisions to negate or validate our assumption that the relationships among different datasets and asset prices are too complex for any human to characterize.

We also noted that our approach is more comprehensive than peers given our meticulousness in ensuring consistency across the forecasting process.. For example, during our class presentations we have not heard from any other team about how they integrated their forecasted prices into the historical price set to analyze and optimize their composition in the portfolio.Many teams also do not update the portfolio optimization algorithm with the latest risk-free rate when determining their optimal weights.

Another area where our approach stood out is how we handled issues with the datasets drawn from Yahoo Finance. For example, in the stock price dataset, ticker OGN has prices missing before 05/14/2021. If anyone was to use this data in an analysis, they have to ensure that they backfill the prices for OGN to ensure consistency across the dataset.

Additionally, cryptocurrencies trade on weekends in contrast to the other 100 assets. When evaluating these assets, we separated them from the overall dataset. After conducting their respective analyses, we concatenate all assets prices on similar dates before their fed into our chosen optimizer.

There were issues with prices for the ETFs as well. Some ETFs have traded on days other ETFs have not or on days where equities have not traded. This has resulted in some prices for ETFs being returned as null values, causing issues in training our model. We made sure to forward fill such cells to ensure consistency and validity in our framework.

Many of our peers have also heavily focused on recreating popular trading strategies or creating trade signals rather than formulating a framework for asset price prediction. Most of such approaches, we believe, are already heavily adopted by market participants and do not provide an edge in identifying best performing assets in an investment universe.

5 Conclusion

Our methodologies for forecasting and optimal portfolio construction fared well against competition and validated our main assumption that computers are more suited in handling asset price prediction than fundamental, traditional approaches to valuation.

We are, however, aware that more work needs to be done in order to make our approach sustainable and comprehensive over the long run.

An overview of other methods we tried incorporating into the algorithm but couldn't due to resource limitations are outlined below:

Scraping tweets on cryptocurrency to create a time series of sentiment to incorporate in the algorithm: couldn't not be used in the model as the mainstream libraries available to extract tweets are slow and furthermore, the results we get can vary widely depending on the keys words we use to identify tweets Training a Convolutional Neural Network to identify price patterns and predict future movement of the asset: training the model proved to be time consuming and onerous given the large amount of screenshots of past price movements we need to feed the model and their variation over different time intervals that can range from a day, a week, fortnight, etc. Extracting market sentiment from news outlets and creating a time series for different asset classes: faced issues with the API for the news outlets we identified.

In conclusion, the project helped us understand and appreciate a few facts about the market. First of them being the importance of diversification. Many of our peers had their focus restricted on a few assets that they believed they had the most capability to predict, but we believe that our move to not restrict ourselves to any particular set of assets helped offset any particularly unfavorable idiosyncratic risks.

Second, we understood the importance securing our access to databases we considered important in our line of work and having plans of action in case we face issues accessing those databases.

Lastly, we saw firsthand the benefit of continuously innovating and updating our approach to the problem statement. Avoiding complacency is key in maintaining an edge over others in the market.

References

A. Menezes, "Stock price simulations," Medium, 29-Jan-2021. [Online]. Available: *https://medium.datadriveninvestor.com/stock-price-simulations-fa2ce492dd93*.

B. G. Teo, "Simulating random walk of stock prices with Monte Carlo simulation in Python," Medium, 19-Sep-2021. [Online]. Available: *https://medium.com/the-handbook-of-coding-in-finance/simulating-random-walk-of-stock-prices-with-monte-carlo-simulation-in-python-6e233d841e*.

Bottama, "Bottama/stochastic-asset-pricing-in-continuous-time: Predicting stock prices using geometric brownian motion and the Monte Carlo Method," GitHub. [Online]. Available: *https://github.com/bottama/stochastic-asset-pricing-in-continuous-time*.

Cboe Global Markets, "Historical Data for Cboe VIX® Index and Other Volatility Indices," Cboe Global Markets. [Online]. Available: *https://www.cboe.com/tradable_products/vix/vix_historical_data/*.

D. Boh, "Easily optimize a stock portfolio using PyPortfolioOpt in python," Medium, 25-Apr-2022. [Online]. Available: *https://medium.datadriveninvestor.com/easily-optimize-a-stock-portfolio-using-pyportfolioopt-in-python-80492b83912a*.

E. Melul, "Monte Carlo simulations for predicting stock prices [python]," Medium, 20-May-2020. [Online]. Available: *https://medium.com/analytics-vidhya/monte-carlo-simulations-for-predicting-stock-prices-python-a64f53585662*.

Federal Reserve Bank of St. Louis, "Effective federal funds rate," FRED, 16-Oct-2022. [Online]. Available: *https://fred.stlouisfed.org/series/EFFR*.

Federal Reserve Bank of St. Louis, "SP 500," FRED, 16-Oct-2022. [Online]. Available: *https://fred.stlouisfed.org/series/SP500*.

Federal Reserve Bank of St. Louis, "5-year, 5-year forward inflation expectation rate," FRED, 16-Oct-2022. [Online]. Available: *https://fred.stlouisfed.org/series/T5YIFR*.

H. Ertan, "CNN-LSTM based models for multiple parallel input and multi-step forecast," Medium, 17-Nov-2021. [Online]. Available: *https://towardsdatascience.com/cnn-lstm-based-models-for-multiple-parallel-input-and-multi-step-forecast-6fe2172f7668*.

Nachi-Hebbar, "Time-series-forecasting-LSTM/rnn_youtube.ipynb at main · Nachi-Hebbar/Time-Series-forecasting-LSTM," GitHub, 19-May-2021. [Online]. Available: *https://github.com/nachi-hebbar/Time-Series-Forecasting-LSTM/blob/main/RNN_Youtube.ipynb*.

Qpm, "Commodity prices dataset (2000 - 2022)," Kaggle, 01-Oct-2022. [Online]. Available: *https://www.kaggle.com/datasets/debashish311601/commodity-prices*.

S. Khankary, "Multivariate Time Series Forecasting using RNN(LSTM)," Medium, 27-Jan-2022. [Online]. Available: *https://medium.com/mlearning-ai/multivariate-time-series-forecasting-using-rnn-lstm-8d840f3f9aa7*.

Time Series Forecasting with RNN(LSTM)| Complete Python tutorial|, YouTube, 19-May-2021. [Online]. Available: *https://www.youtube.com/watch?v=S8tpSG6Q2H0amp;list=PLqYFiz7NM$_S$MC4ZgXplbreXlRY4Jf4zBPamp;index* = 11.

U. Singhal, "Building an optimal portfolio with python," Trade With Python, 21-Jul-2021. [Online]. Available: https://tradewithpython.com/building-an-optimal-portfolio-with-python.

U. Singhal, "Portfolio analysis using Python," Trade With Python, 05-Jul-2021. [Online]. Available:

https://tradewithpython.com/portfolio-analysis-using-python.

V. Lendave, "How to do Multivariate time series forecasting using LSTM," Analytics India Magazine, 09-Jul-2021. [Online]. Available: https://analyticsindiamag.com/how-to-do-multivariate-time-series-forecasting-using-lstm/.