Portfolio Risk Assessment and Rebalancing using Machine Learning

Kiran, Nandita Dwivedi and Muhammad Rafay Aleem

1. List 3 questions that you intend to answer (1 point)

1. Can we find the ideal risk level for a diversified investment portfolio composed of equity holdings?
2. Can we identify the factors affecting prices of ETF (Exchange Traded Funds) using historical stock market data and corresponding SEC (Securities and Exchange Commission) filings?
3. Can we maximize profits by dynamically rebalancing fund allocation for portfolio holdings?

2. List all the datasets you intend to use (1 point)

1. SEC's EDGAR (Electronic Data Gathering, Analysis, and Retrieval system)
2. Quandl API
3. Financial news dataset from Reuters
4. New York Times Article Search API
5. Edgar DataFied API

3. Give us a rough idea on how you plan to use the datasets to answer these questions. (2 points)

- Data Collection:
  The above listed datasets will be extracted using available APIs and scraped. We can use crawling frameworks to collect the SEC filings (10-K, 10-Q, 8-K filings) of various companies.

- Data Exploration:
  EDA needs to be conducted to understand the data better and identify relationships and trends between stock prices, sentiments, etc. For example, we can use EDA to identify any relationship between fluctuations in the market and emotions gathered from SEC 10-Q filings. We can also explore if certain ETFs tend to perform better at certain time intervals which can identify market seasonalities.

- Data Cleaning:
  The above data sources might require cleaning in order to integrate different sources or eliminate the information that is not needed for the analysis, this can considerably

reduce the size of the data and make the processing easier. We can use pyspark for the cleaning process.

- Data Integration:
  We will be using data from multiple resources. Since stock prices, financial news and SEC filings data are all non-overlapping sets, we will need to stitch them together to enable analysis and find correlations. For example, predicting market movements for the next month could be based on sentiments obtained from previous quarters SEC 10-Q filing.

- Data Analysis:
  We can explore basic statistic models like logistic regression for binary analysis, even models like SVM can prove to be useful. Natural Language Processing will be needed for sentiment analysis on SEC's EDGAR and New York Times articles. Python libraries like Theano, scikit-learn and Pandas can be useful for the analysis. The results can be evaluated by understanding the correlation between the sentiments, performing regression analysis and measuring the statistical measurements like accuracy and variation explained.

- Data Product:
  Our final product will be a combination of Jupyter notebooks, various visualization and a web page comprising the findings (best risk level and best possible allocations) for an entire year.

4. Think about that once your project is complete, what impacts it can make. Pick up the greatest one and write it down. (1 point)

Optimizing equity investments is the major motivation for any investor, whether it's an individual or an investment firm. While mathematics hasn't been able to beat the markets so far, we intend to build a solution that is pragmatic for any investment firm focused on maximizing profits while mitigating risks. This work can be impactful for firms which manage portfolios on behalf of their clients. They can leverage our work to dynamically allocate equity holdings within a certain risk bracket and maximize returns for that time period.