Final Project: Proposal

Anurag Bejju, Manan Parasher, Rishab Singh, Nikitha Ravi

1. List 3 <u>questions</u> that you intend to answer (1 point)

- Can we find trends that showcase a correlation between Stocks and Crypto-currency markets?
- Can a multi-purpose model be designed that forecasts market trends for Stocks and Cryptocurrencies alike based on intersecting OHLC features?
- To what extent does global factors (captured by daily news) influence the volatility witnessed in Stock and Crypto Market?

2. List <u>all the datasets</u> you intend to use (1 point)

Stock and *Cryptocurrency* historical data with hourly price fluctuations for the last few years will be used. In addition to that statistics from each cryptocurrency's *GitHub repo*, textual information from *discussion forums* and social media metrics from *Twitter* will be collected and used. Also financial news for the last one year will be collected using *Financial Times* API to get contextual understanding of each trade happened.

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

- **<u>Data Collection:</u>** Where/how to get data

  The data used for a model is a compiled from 4 different sources.
  - OHLC hourly financial data for Cryptocurrency are collected by web scraping sites like *CoinMarketCap.com* and for stocks its compiled using free API's provided by *iextrading.com*. Also some textual and statistical information is also scraped from their individual *GitHub repo*.
  - *Twitter API* will be used to collect social media quantitative metrics *Financial Times API* to get contextual understanding of the trade for each cryptocurrency and stock being used.
  - All this data will also be collected in real-time to support our hourly ETL jobs

  Tech Stack:
  - ○ *Data Collection:* BeautifulSoup, Request, lxml, python-web scraper etc
  - ○ *Data Storage:* SQLite3, MySql, Cassandra, Hive etc.

- Data Exploration: Do you need to conduct EDA in order to understand the data?

  - Since the data is being collected from multiple sources and is not properly correlated with each other, EDA will be used to access what information is useful and what information can be discarded. It will also be used to analyze features that have an impact in determining the volatility witnessed in Stocks or Cryptocurrency markets.
  - EDA will be continuously incorporated throughout the project development process to make our model as compact and precise as possible.

  Tech Stack: Matplotlib, pandas, plotly

- Data Cleaning: Do you need to clean data? How to clean them?

  - Since the data collected through API's and web scraping websites will be in different formats and might not be in the form we need, we will have to perform some amount of ETL work to make it more suited based on our needs.
  - The news feed collected from Financial Times has to undergo some transformation to get a collection of words that are of significant importance.
  - Some API's return information in JSON format that needs to be properly parsed and appended to our asset table

- Data Integration: Do you need to integrate data from multiple sources?
  - Yes. Web scraped data from GitHub and CoinMarketCap.com as well as API data from Twitter, Financial Times and iextrading.com needs to be combined to form our feature vector for both our financial assets (i.e. Cryptocurrency and Stocks)

- Data Analysis: What analysis do you intend to do? (e.g., SQL, Statistics, Deep Learning) How to evaluate your analysis results? (e.g., evaluation metrics, confidence intervals, benchmark)

  - On a broader scale, we intend to do *statistical analysis* that provides information such as top performing financial asset, day wise performance etc.
  - We also perform *Natural Language Processing* operations like sentiment analysis, Jaccard Similarity, key word extraction, TF-IDF, Word2Vec on our textual data collected from twitter, discussion forums and news from financial times.

- *Machine learning* will be used to train our Regression or Classification Model that can effectively predict market fluctuations and price for each financial asset based on our input feature vector.
- Finally *evaluation metrics* will be used to test our accuracy and performance of our trained model.

- ***Tech Stack:*** PyTorch or PySpark

- Data Product: What product do you want to build? (e.g., visualizations, an interactive web app, a jupyter notebook)

  We intend to develop and design an interactive web application with multiple sections/pages with intuitive visualizations. Some of the widgets in the dashboard may contain:
  - *Historical Data Aggregation:* Top Performing Stocks per day, Gainer/Loser chart, Time-Series Price Ticker Chart, Volume Traded per day etc.
  - *Assets Information (Stocks and Cryptocurrencies):* Financial Asset Information like Asset Name, Asset ID, Symbol, Sector etc.
  - *Global Factors that influence asset price:* Price Predictors based on latest news, Hourly Sell/ Buy Predictors for each asset etc.
  - *Domain Specific Asset Information:* Hourly Performance of each sector, Gains and Losses for each sector etc.

Tech Stack:

- *Front-End:* Flask, JavaScript, HTML, CSS, D3.js, Tableau
- *Real-Time Job Scheduler :* Celery, PySpark, RabbitMQ
- *Hosted on:* SFU Cloud

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)

Based on an article published in *Financial Times*, profits for financial asset managers rose to *$102bn* making it a lucrative market to be in. With more and more people diversifying their asset portfolio and with the crypto market projected to reach *US$6702.1 mn* in 2025, our product intends to helps investors make informed decisions while buying or selling stocks or cryptocurrencies. We intend to target both the traditional and exploratory traders by providing a robust application that can help them make data-driven decisions as well as actively support novice traders by providing intuitive financial predictions based on historical and contextual information in real-time.