Chris Toomey, Jenice An, and Tongyu Guo

ADAN 8888 - Applied Analytics Project - Spring 2026

Module 2 Assignment - Ingest and Explore the Dataset

HelpHerInvest: A machine learning-based investment recommendation system that aligns user interests and risk preferences with benchmark stock evaluation.

Our team's project is to create a recommendation system which develops a personalized stock portfolio based on a user's inputs, such as areas of interest, and optimizing returns. The recommendation system is intended to use a multinomial classification model to determine which stocks are optimal in terms of returns based on the user's inputs.

The first portion of this system is to utilize natural language processing ("NLP") to rank the stock tickers on similarity based on the user's inputs and variables such as the sector, industry, and business summary. The NLP algorithm will then compute a similarity score (such as cosine similarity) between the embedded user inputs compared against the embedded stock representations (sector, industry, business summary). The dataset contains approximately 10,200 rows and 224 columns of data. This data ranges from string and numeric data types, but the variables being focused on within this dataset are of the string data type.

- **Sector**: Sector represents a specific area of the economy the business operates in. There are 11 main sectors such as health care, energy, technology, etc.
- **Industry**: Industry represents a sub-segment of the sector the business operates in. This provides a more detailed area in which the business operates. For example, if a company fell under the technology sector, their industry may be semiconductors (i.e. Nvidia) or consumer-electronics (i.e Apple)
- **Business Summary**: The business summary contains text, roughly the size of a paragraph, providing more details about the business. This field could contain information such as the different business channels it operates in, the products / services offered, where they operate, the company headquarters, etc.

The output will be approximately the top ~50 tickers that are most similar based on the user's inputs. Given that the target variable is unlabeled and we do not currently have a validation dataset, this would fall under the category of unsupervised semantic retrieval (Van Gysel). If our team does require and think it is most optimal to create a supervised learning algorithm to extract the most similar stock tickers, we will create a dataset that maps labels to tickers to validate the user inputs against the output results. Initial research suggests that the unsupervised semantic retrieval approach can be just as effective, if not more effective, than developing a supervised learning model or labeled data.

The second portion of the stock portfolio recommendation system is to develop a multinomial classification model based on the subset of tickers produced from the NLP output. The subset of tickers will be used to perform technical analysis based on changes in historical data. The target variable in this model is the forward excess return, which is calculated by subtracting the return of stock minus a benchmark over a defined period of time. The benchmark we decided to use is the S&P 500, or SPY Exchange-Traded Fund (“ETF”), because it is a good metric to represent the state of the United States economy and baseline return to compare an individual stock against. Our team found the forward excess return to be a suitable target variable because it is a useful metric to compare against stocks across different sectors and sizes.

Forward Excess Return = Return of Stock $_{t+h}$ *- Return of Benchmark* $_{t+h}$
t = Current Time
h = Forecast Horizon (i.e. 3 months, 12 months, etc.)

The predictor variables we are currently working with are degrees of momentum (1-month, 3-month, 6-month,12-month, and 12-1 month), relative strengths against the benchmark, periods of volatility, drawdown metrics, and a moving average. Each of these metrics are based on the last day of the trading month and each row in the dataset represents a ticker. This dataset is calculated and constructed by using the historical data taken from the yfinance API. The size of this dataset is based upon how many tickers are being used and the number of days being analyzed, but it is estimated to be 8500 rows and 15 columns.

Example

Date	**Ticker**	**X_1**	**X_2**	**…**	**X_n**
2026-01-30	AAPL	0.25	0.42	…	0.36

- **1-Month Momentum**: The price percentage change over the last month
- **3-Month Momentum**: The price percentage change over the last 3 months
- **6-Month Momentum**: The price percentage change over the last 6 months
- **12-Month Momentum**: The price percentage change over the last 12 months
- **12-1-Month Momentum**: The price percentage change over the last 12 months, excluding the current month

- **3-Month Relative Strength Against S&P:** 3-month momentum difference from the S&P 3-month momentum
- **6-Month Relative Strength Against S&P:** 6-month momentum difference from the S&P 6-month momentum
- **12-Month Relative Strength Against S&P**: 12-month momentum difference from the S&P 12-month momentum

- **3-Month Volatility**: 3-month volatility calculation based on the standard deviation
- **6-Month Volatility**: 6-month volatility calculation based on the standard deviation

- **6-Month Drawdown**: 6-month comparison maximum stock price against current price
- **12-Month Drawdown:** 12-month comparison maximum stock price against current price

- **200-day rolling average:** 200-day rolling average of the stock price
- **Percentage above 200-day rolling average**: The current stock prices percentage above/below the 200-day rolling average price

References:

Van Gysel, C., de Rijke, M., & Worring, M. (2016). *Unsupervised, efficient and semantic expertise retrieval*. arXiv. https://doi.org/10.48550/arXiv.1608.06651