INST737 - Introduction to Data Science

Milestone - 1 Report

1a.Research Questions

1. Can historical stock price data be used to predict the future stock prices of FAANG companies?
2. Could the unemployment rate and the nation's GDP impact the overall stock market performance of FAANG companies?

In the world of finance, we're diving into two fascinating research questions. The first one is all about trying to predict what the future holds for the stock prices of major players like the FAANG companies (Facebook, Apple, Amazon, Netflix, Google). We're digging into historical stock prices to see if we can figure out where these stocks might be headed next. This research can help investors make smarter decisions, fuel high-tech trading strategies, assist portfolio managers, give companies a handle on stock market risks, and even guide governments in shaping economic policies.

The second question explores the fascinating connection between a country's unemployment rate and its Gross Domestic Product (GDP) and how these two factors can sway the overall performance of FAANG companies in the stock market. Moreover, this research isn't just academic; it's got real-world implications. It can help with economic predictions, influence government policies, and even make risk management strategies smarter. In essence, it helps regular folks, businesses, and policymakers navigate the financial world with more insight and confidence.

1b. State of the Art

To explore the various factors that impact the stock market, we have selected four papers that investigate the relationship between stock market, GDP, and the unemployment rate, as well as the utilization of machine learning techniques in stock market forecasting.

- **Stock prices and GDP in the long run (Alexius & Spång, 2018)**

  This paper explores the long-run equilibrium relationship between stock prices and GDP in the G7 nations. Their research indicates that both domestic and foreign GDP have a significant impact on stock prices, offering evidence of long-term equilibrium relationships that go beyond typical findings.

- **The Impact of Inflation, GDP, Unemployment, and Money Supply On Stock Prices (Shiblee, 2009)**

  The research focuses on the New York exchange and analyzes data from 1994 to 2007, and argues that money supply is the most important influencing factor on stock price, outweighing the impacts of inflation and unemployment.

- **The Stock Market/Unemployment Relationship in USA, China and Japan (Farsio & Fazel, 2013)**

  Contrary to common view points, the authors challenge that unemployment rates do not have a causal relationship with stock prices. By analyzing quarterly data from the USA, China, and Japan over the 1970-2011 period, the authors caution against relying on unemployment data for investment decisions.

- **Machine learning techniques and data for stock market forecasting: A literature review - ScienceDirect (Kumbure et al., 2022)**

  This article provides a general review on the data and machine learning methods used in stock market prediction, assessing 138 journal articles published from 2000 to 2019. The study revealed that supervised learning methods are commonly used in forecasting, and that combining different machine learning methods has proven effective in improving prediction accuracy.

The four papers we reviewed provide various views of the factors affecting stock market performance, as well as the techniques used in the field. Our research will contribute to the existing knowledge by focusing on specific variables and their impact on FAANG companies. By doing so, we aim to provide insights with practical implications that extend beyond academia, assisting a variety of stakeholders in making more informed financial decisions.

1c. Datasets

In this project, we selected the following three datasets:

- Historical Stock Price of (FAANG + 5) companies
- Unemployment Rate
- US Monthly GDP (MGDP) Index

- **Historical Stock Price of FAANG companies**
  The Historical Stock Price dataset encompasses historical stock price data for FAANG companies, featuring essential columns such as date, opening price, highest price, lowest price, closing price, adjusted closing price, and trading volume. General statistical analysis is conducted on the continuous numerical columns such as opening price,

highest price, lowest price, closing price, adjusted closing price, and trading volume to gain insights into central tendencies and variability. Dataset includes numerical (in dollars), date and text variables with 8,805 rows and 7 columns.

- **US Monthly GDP (MGDP) Index**

  The dataset comprises two columns of monthly economic data spanning the years 1992 and 1993, with each row corresponding to a specific month. It includes two key variables: the Monthly Nominal GDP Index, representing the unadjusted economic output value for goods and services produced within a country each month, and the Monthly Real GDP Index, which adjusts for inflation to offer a more precise indicator of economic growth. General statistics such as mean, median, standard deviation, minimum, and maximum values are computed for both Nominal and Real GDP. Dataset includes numerical variable (dollars), date variables with 378 rows and 3 columns. This dataset holds significance for economic research and analysis.

- **Unemployment Rate**
  The dataset from the Federal Reserve Economic Data (FRED) focuses on the Unemployment Rate in the United States, offering a monthly, seasonally adjusted perspective. The "UNRATE" column signifies the Unemployment Rate, measured in percentages, reflecting the proportion of the labor force actively seeking employment in a given month. Mean, median, standard deviation, minimum, and maximum is calculated on UNRATE column (Unemployment Rate). Dataset includes numerical variable (percentage), date variables with 908 rows and 2 columns This dataset serves as a valuable resource for comprehending historical employment patterns and labor market conditions in the United States.

1d. Data Cleaning Efforts

We have executed a sequence of data processing and analysis tasks in R. Initially, we loaded and merged several CSV files from Amazon, Google, Facebook, Apple, and Netflix, each containing columns: "Date," "Open," "High," "Low," "Close," and "Adj.Close." We filtered the rows to include only those with dates from January 1st, 2015, onwards. To differentiate between the datasets, a "Source" column was added to the merged data frame, and we rearranged the columns to position "Source" as the first column. After the merging, we checked the entire dataset for any null values and found none. The goal throughout this process was to refine and prepare the data for subsequent analysis.

1e. Other Software Engineering Efforts

In this project, we used R Studio, Excel, and Python to address different aspects of the data science pipeline and answer our research questions. R studio was used for data cleaning, analyses and visualization, where specialized packages offer an efficient way to interpret our data sets. Excel served as an easy-to-use starting point for inspecting data and setting the groundwork for further analysis. For more complex tasks, like merging the US Monthly GDP Index with Unemployment Rate datasets, we deployed Python scripts that utilized the Pandas library for column standardization and data merging.Together, these combined software engineering and data collection efforts streamlined our workflow, enabling us to answer our research questions comprehensively.

Contribution Section

Question 1a(Research Questions): developed by Srikanth Parvathala;

Question 1b(State of the Art): developed by Ya-Ting Yang;

Question 1c(Datasets): developed by Srikanth Parvathala;

Question 1d(Data Cleaning Efforts): developed by Vijay Arni;

Question 1e(Other Software Engineering Efforts): developed by Ya-Ting Yang;

Vijay Arni prepared 100% of the R code;

Ya-Ting Yang prepared 100% of the Python code;

All members contributed equally to the preparation and recording of the presentation;

Reference

Alexius, A. (2018). *Stock prices and GDP in the long run*. https://econpapers.repec.org/article/sptapfiba/v_3a8_3ay_3a2018_3ai_3a4_3af_3a8_5f4_5f7.htm

Farsio, F., & Fazel, S. (2013). *The Stock Market/Unemployment relationship in USA, China and Japan. International Journal of Economics and Finance, 5(3)*. https://doi.org/10.5539/ijef.v5n3p24

Historical Stock Price of (FAANG + 5) companies. (2021, December 30). Kaggle. https://www.kaggle.com/datasets/suddharshan/historical-stock-price-of-10-popular-companies/code?select=Microsoft.csv

Kumbure, M. M., Lohrmann, C., Luukka, P., & Porras, J. (2022). Machine learning techniques and data for stock market forecasting: A literature review. *Expert Systems With Applications*, *197*, 116659. https://doi.org/10.1016/j.eswa.2022.116659

Sharing insights elevates their impact. (n.d.). S&P Global. https://www.spglobal.com/marketintelligence/en/mi/products/us-monthly-gdp-index.html

Shiblee, L. S. (2009). The Impact of Inflation, GDP, Unemployment, and Money Supply On Stock Prices. *Social Science Research Network*. https://doi.org/10.2139/ssrn.1529254

Unemployment rate. (2023, September 1). https://fred.stlouisfed.org/series/UNRATE