Macro Indicators & Web-Scraped Market Sentiment Analysis for NASDAQ Trends

Final Report – DSCI 510 Principles of Programming for Data Science

Student: Shiyi Wang

USC ID: 9862305589

Email: shiyiw@usc.edu

1. Project Description

This project investigates how **macro-economic indicators** (e.g., GDP, CPI, FEDFUNDS, retail sales) and **web-scraped financial news sentiment** relate to the behavior of the **NASDAQ Composite Index**.

The project pursues three main goals:

1. **Explanatory Analysis:**
   Assess long-term correlations between macroeconomic indicators and NASDAQ movements.
2. **Predictive Modeling:**
   Evaluate whether macro variables can classify **60-day NASDAQ tail-risk events** (defined as future drawdowns ≤ −2%).
3. **Sentiment Analysis:**

   real-time financial news headlines were scraped and analyzed using sentiment scoring to understand short-term market tone

2. Data

2.1 Data Sources

In this project, I collected data from **four main sources**, using a combination of **APIs, web scraping**, and **market data retrieval tools**:

1. **FRED API (Federal Reserve Economic Data)**
   - Accessed via the official FRED REST API.
   - Collected 11 macroeconomic and financial indicators:
     GDP, CPIAUCSL, UNRATE, FEDFUNDS, INDPRO, RSAFS, HOUST, DGS3MO, VIXCLS, STLFSI, TEDRATE.
2. **Yahoo Finance API via yfinance (Market Index Data)**
   - Retrieved **Wilshire 5000 Index (^W5000)** closing levels.
   - Retrieved **NASDAQ Composite Index (^IXIC)** closing levels.
3. **Yahoo Finance News Search Endpoint (Scraped JSON)**
   - Queried news using keywords such as *"nasdaq"*, *"nasdaq futures"*, *"stock market"*, etc.

- Extracted article titles, links, publish timestamps, and publishers.

4. **Local Script-Based Data Processing**
   - All raw datasets were saved into the directory:
     data/raw/
   - Each dataset was retrieved consistently from the Python script above for reproducibility.

2.2 Number of Data Samples Collected

Dataset	Description	Number of Observations
US Macroeconomic Indicators (FRED)	11 time series from 2014–12–31 to 2025-12-9 (newest)	~120 monthly observations × 11 series
Wilshire 5000 Index	Daily close prices	~2,895 daily observations
NASDAQ Composite Index	Daily close prices	~2,895 daily observations
Yahoo Finance News	News articles for 5 search queries	~50 articles (10 per query)

✓ Total structured time-series data: ~2,895 rows
✓ Total news articles collected: ~50 records

3. Data Cleaning, Analysis & Visualization

3.1 Data Cleaning

All cleaning steps were implemented through clean_data.py.

Macroeconomic Data

- Converted FRED API JSON responses into structured DataFrames
- Standardized all timestamps and set them as indices
- **Forward-filled missing** macro values (common for monthly indicators)
- Joined all macro series into one unified table
- Loaded Wilshire 5000 and forward-filled holidays / missing entries to align frequency
- Constructed the **Buffett Indicator**:

$$\text{Buffett Indicator} = \frac{\text{Wilshire 5000}}{\text{GDP}}$$

- The cleaned macro+ Buffett Indicator dataset was merged with the NASDAQ close series, missing values were also forward-filled.
- The Wilshire 5000 column was dropped after the Buffett Indicator was computed because the raw index is no longer needed in the processed dataset.
- To ensure the final dataset is suitable for modeling, columns with **no data available after 2025** were removed
- The final macro-market dataset was saved to: **data/processed/processed_market_data.xlsx**

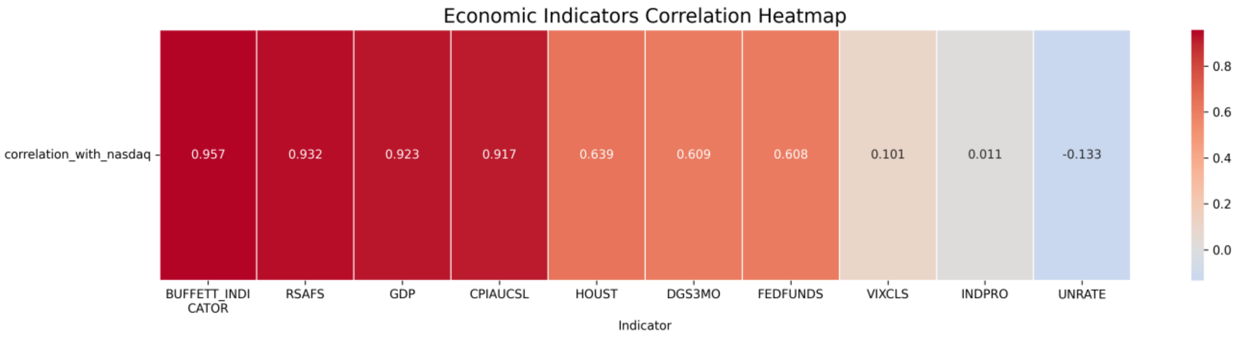
News Data

- Parsed raw API responses
- Converted UNIX timestamps to **readable datetime formats**
- **Removed duplicate** headlines
- Extracted clean fields: title, publisher, date, and source
- **Index was reset** to a sequential integer index
- The final cleaned news dataset was exported to: **data/processed/processed_yahoo_news.xlsx**

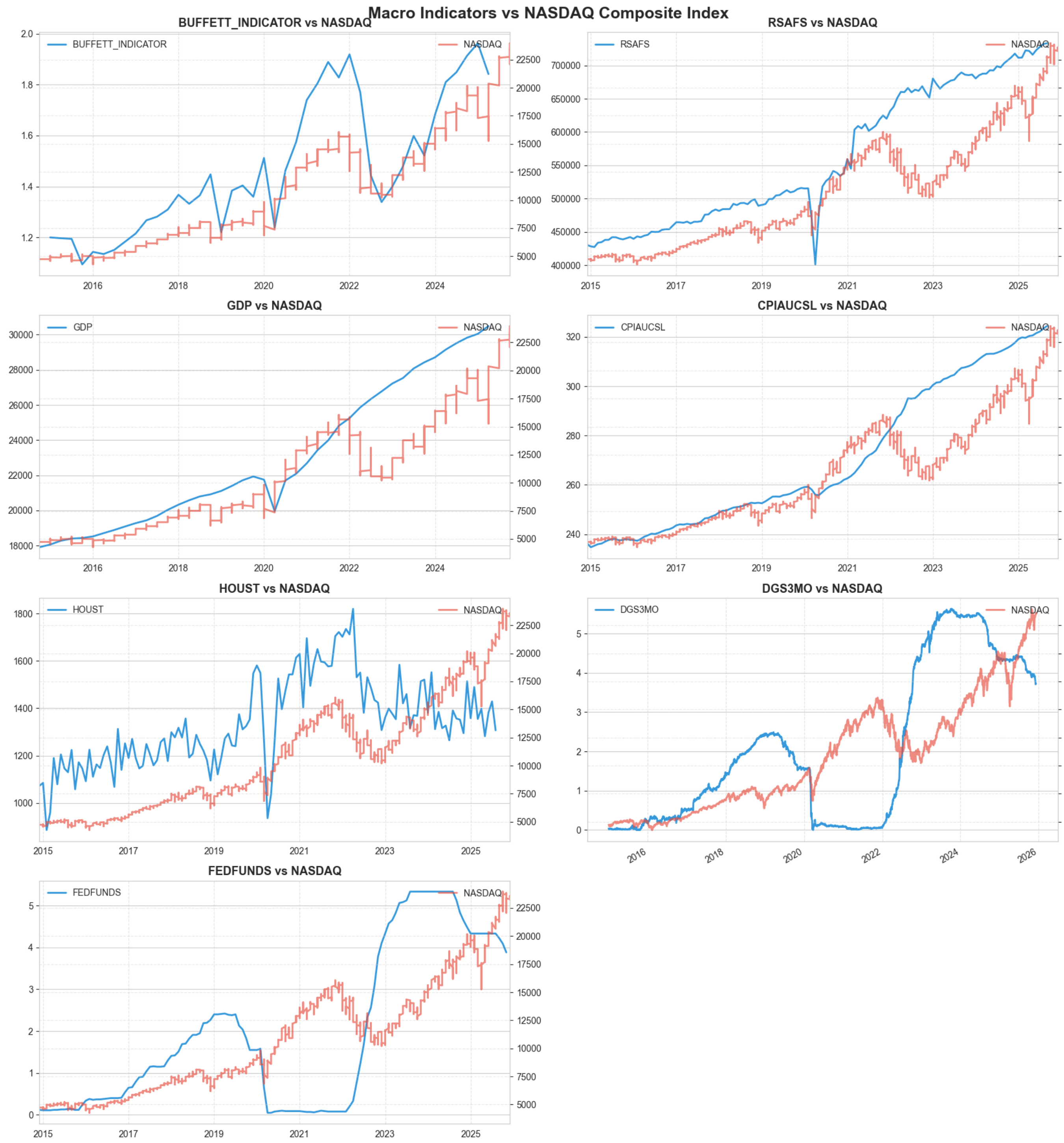
3.2 Data analysis and Visualization

3.2.1 Correlation Analysis and Visualization Between Macro Indicators and NASDAQ

Using the cleaned dataset, I calculated Pearson correlations between each macroeconomic indicator and the nasdaq_close level. As shown in the legend below:



Macro variables exhibited mixed correlation strength with NASDAQ prices. Indicators such as liquidity conditions, valuation (Buffett Indicator), and economic growth showed the strongest directional relationships. I identified “strong features” (|correlation| > 0.5) for potential inclusion in further modeling, as shown in the Macro–NASDAQ time-series panel below:



3.2.2 Tail-Risk Modeling Using Logistic Regression

The main predictive task in the project is estimating whether the NASDAQ will experience a **tail-risk event.**

Tail-risk defined as:

$$\text{TailRisk} = \begin{cases} 1 & \text{if drawdown over next 60 days } \leq -2\% \\ 0 & \text{otherwise} \end{cases}$$

Steps:

1. Calculated 60-day forward minimum price
2. Computed percentage drawdown

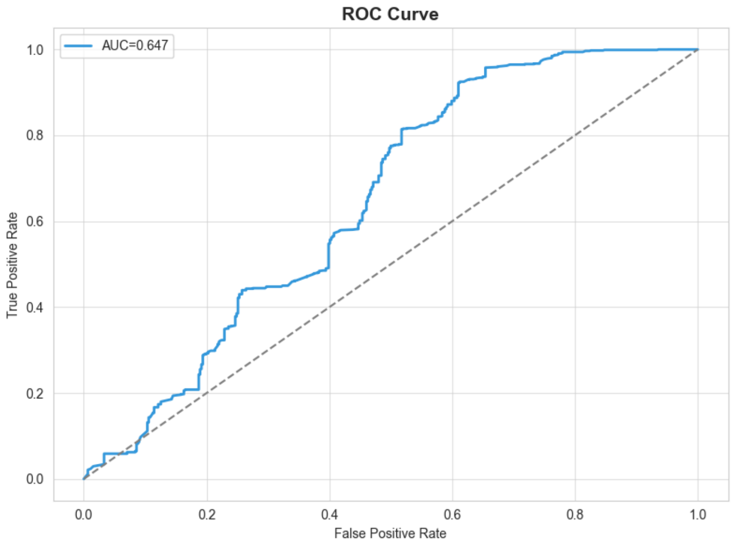
3. Labeled samples as high-risk vs. stable

4. Train/test split at 2021-01-01

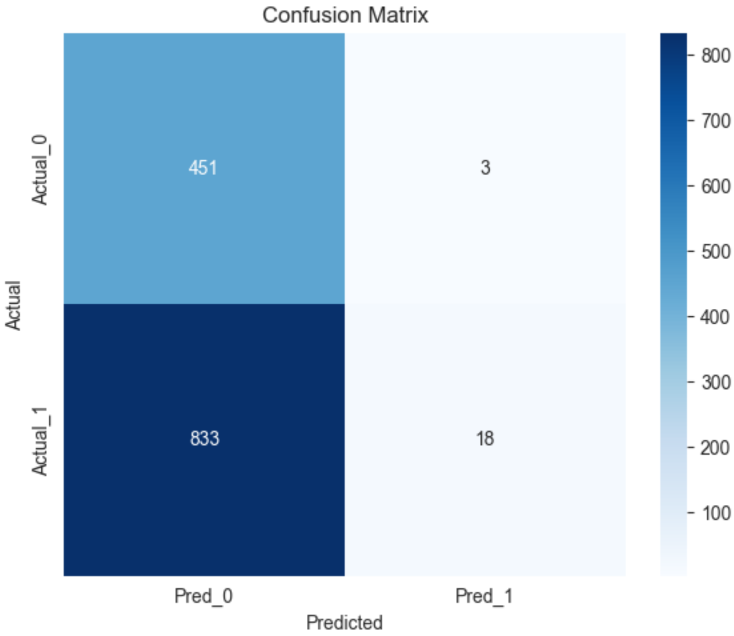
5. Trained Logistic Regression model, input features: BUFFETT_INDICATOR, RSAFS, GDP, CPIAUCSL, HOUST, DGS3MO, FEDFUNDS

Model Performance:

The model achieved: AUC = 0.647, as shown in the below legend. AUC > 0.5 suggests some useful signal, but performance is far from strong due to the difficulty of predicting tail events.



The classification report revealed an extreme class imbalance, as shown in the confusion matrix and table below:



Class	Precision	Recall	F1	Support
0 (no tail-risk)	0.35	0.99	0.52	454
1 (tail-risk)	0.86	0.02	0.04	851

Conclusion:

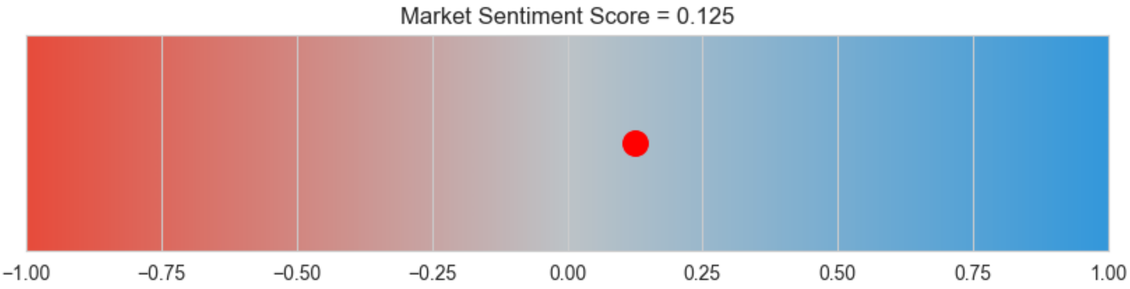
While macroeconomic indicators show some correlation with equity trends, they provide very limited predictive power for short-term tail events. Predicting 60-day future crashes using monthly macro data is inherently challenging.

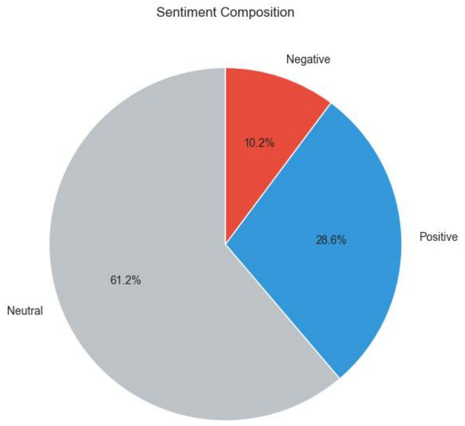
3.2.3 Market Sentiment Analysis Using News Headlines

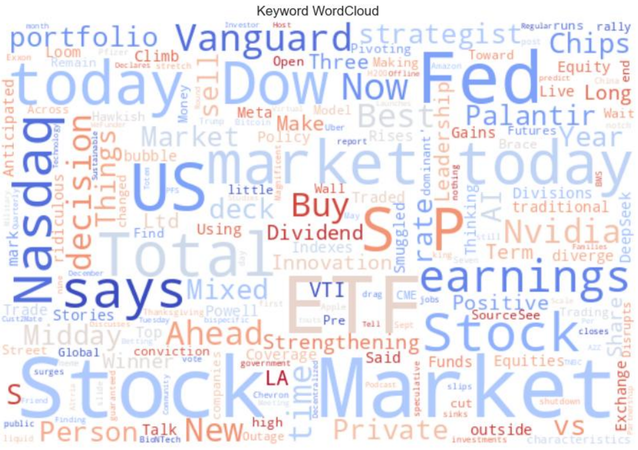
To incorporate qualitative market information, I performed sentiment scoring on Yahoo Finance news headlines using the VADER sentiment analyzer.

- Generated sentiment distribution (positive/neutral/negative)
- Computed market sentiment score:
  Average sentiment = 0.125 (*dynamic and changes with latest news*)

As shown in the legend below:







3.3 Hypothesis and Conclusions

Hypothesis

1. **Macro indicators are correlated with long-term NASDAQ movements.**
2. **These indicators can provide moderate predictive ability for tail-risk events.**
3. **Short-term sentiment may complement macro signals.**

Conclusions

- Several macro indicators do show **strong correlation** with NASDAQ levels, supporting Hypothesis 1.

- However, Logistic Regression results show **poor recall** for tail-risk prediction, meaning macro indicators **alone are insufficient** for reliable crash prediction (Hypothesis 2 partially rejected).
- News sentiment provides **useful contextual information** (e.g., negative spikes during volatility), supporting Hypothesis 3, but its predictive usefulness was not formally tested.

Overall, macroeconomic variables help explain trends but **do not reliably predict short-term tail-risk**, aligning with financial theory that macro signals move slowly compared to market volatility.

4. Changes from Original Proposal

Original Plan	Final Implementation	Reason
Include more advanced ML models (Random Forest, XGBoost)	Reduced to Logistic Regression only	Course guidelines emphasize simplicity and interpretability
Predict individual stock drawdowns	Shifted to NASDAQ index only	Data volume & runtime constraints

5. Future Work

If more time or computational resources were available, future extensions include:

- Implementing **Random Forest, XGBoost, or LSTM** for improved predictive power
- Incorporating **macro regime detection** (e.g., recession vs. expansion)
- Collecting **longer sentiment time-series** instead of latest 50 headlines and multiple news resources
- Testing lagged macro effects (lead/lag correlations), explore some trigger events.