Executive Summary: Stock Insights from Global News Sentiments (SIGNS)

Data Science Boot Camp The Erdős Institute - Fall 2024

Team Members: Rishabh Bhardwaj, Yaman Sanghavi, Kunal Mozumdar, Hirak Bandyopadhyay

Github: https://github.com/hirakban/SIGNS_2024.git

Research Questions and Objectives:

- **Objective**: Assess the impact of financial news and global news sentiment on the stock price variability of five major US tech stocks: Google, Microsoft, Amazon, Apple, and NVIDIA.
- **Data Collection**: Gather news data from diverse sources, focusing on topics like financial markets, geopolitics, climate change, global conflicts, and major election results.
- **Analysis**: Evaluate the influence of these news sentiments on the predictability of stock price trends.
- **Optimization**: Utilize predictive models to optimize stock portfolios for superior performance compared to baseline models such as ARIMA or Buy-and-Hold strategies.

Data gathering and Feature selection:

- **Data Collection**:
  - **Stock Price Data**: Retrieved historical financial metrics from Yahoo Finance.
  - **News Data**:
    - Developed a news scraper to collect headlines from Reddit, Google, and News API.
    - Augmented global news data with a Kaggle dataset and additional recent news scraping.
- **Sentiment Analysis**:
  - Used **VADER** to extract sentiment scores for financial news.
  - Developed a custom lexicon to compute sentiment scores for global news.
- **Exploratory Data Analysis (EDA)**:
  - Analyzed correlations between stock variables and financial news sentiment scores.
  - Incorporated technical analysis tools into the modeling process.
- **Modeling and Testing**:
  - Tested regression and classification models to identify the most significant stock value indicators.
  - Evaluated model accuracy in predicting stock price trends.

Model building and Validation:

- **Baseline Model**:
  - Implemented an **ARIMA model** with news sentiment scores as exogenous regressors to capture historical trends and forecast stock prices.
- **Machine Learning Models**:
  - **Target Variables:** Stock Moving Average (SMA) and Closing Price Difference
  - **Classification**: Logistic Regression, Gradient-Boosted Trees, RandomForrest and XGBoost.
  - **Regression**: Multilinear Regression, Gradient-Boosted Trees, RandomForrest and XGBoost.
- **Model Training**: Trained on 80% of the dataset and tested on the remaining 20%.
- **Portfolio Simulation**:
  - Simulated capital gain scenarios for stocks within a limited investment window.
  - Trained the models on a timeline of November 2020 to October 2023 and ran the portfolio simulation over May 2024 to November 2024 timeline.

Results and Future Directions:

- **ARIMAX** accurately captures historical SMA trends but struggles with future predictions.
- **XGBoost** outperforms other models, improving accuracy by 2% with sentiment scores.
- **XGBoost** and **Gradient-Boosted Trees** with global news sentiment scores deliver returns of 25% and 30% respectively, outperforming traditional valuation models such as Buy-and-Hold strategies (12%) .
- **Conclusion**:
  - Financial and global news sentiment scores are critical for predicting financial indicators and optimizing portfolios.
  - Models leveraging sentiment data significantly enhance investment returns and reduce losses under stress-testing.
- **Future Directions**:
  - Extend the methodology to other S&P 500 stocks and sectors.
  - Explore applications for risk assessment and ETF investment strategies.
  - Develop advanced models using deep neural networks (DNNs).
  - Improve the portfolio strategy by introducing a robust risk management strategy that takes into account factors like market volatility.
  - Use deep learning and transformer based sentiment analysis tools in Python like BERT, RoBERTa, or DistilBERT for more robust feature engineering.