# CSE6242 - Final Project - Fall 23 - Team051

Aniket P Srivastava, Rajiv Kamal, Longhao Wang, Inderjeet Singh, Amrapali Shyam, and Umesh Jadhav

## Predicting S&P 500 Index and Correlation with Drivers

Made using Canva

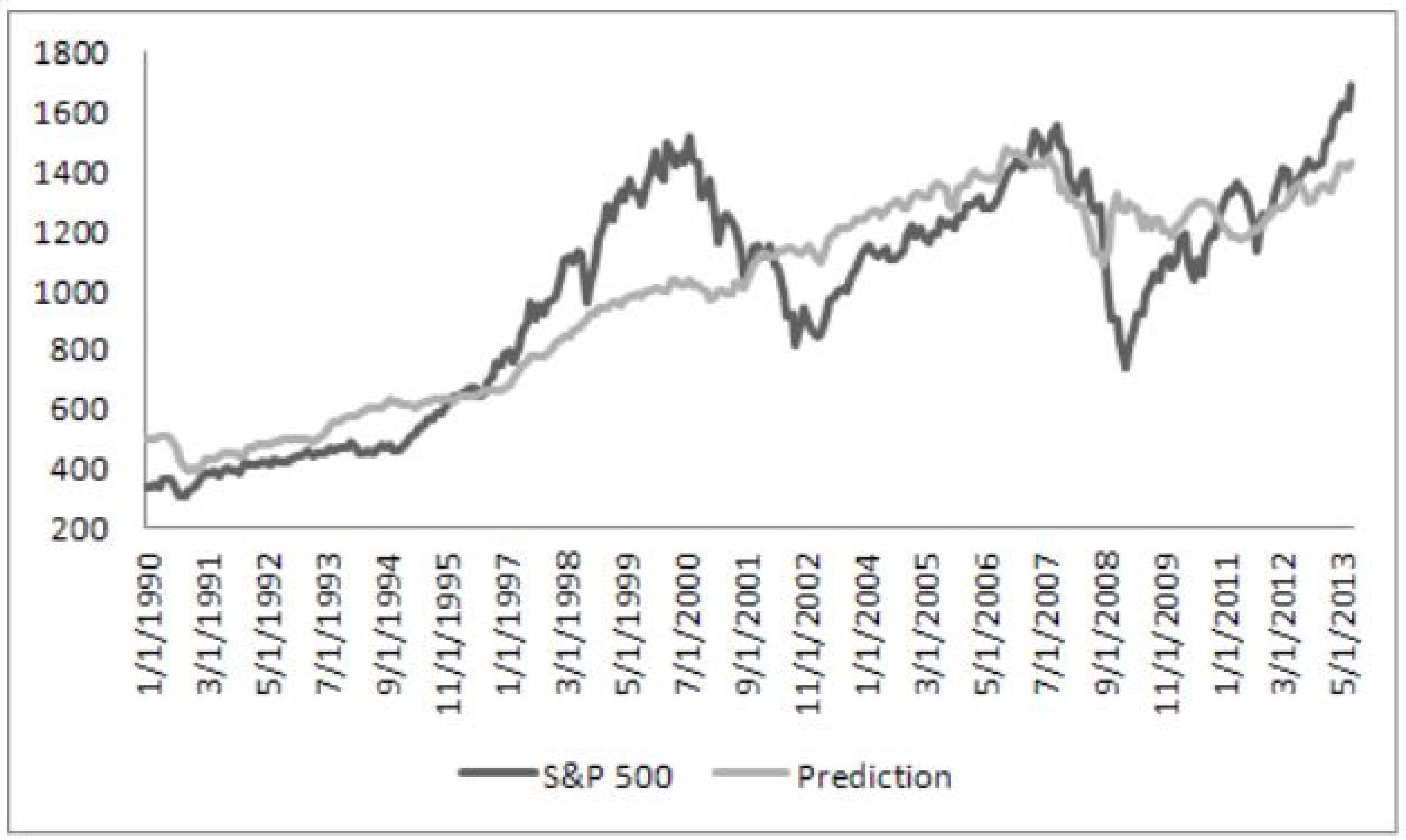
## Motivation/Introduction

### The problem

- Improving the precision of predictive models for the globally significant S&P500 index.
- Employing an ensemble of advanced methodologies like GBM, Random Forest and Recurrent Neural Networks (RNNs) to achieve higher levels of accuracy in forecasting.
- Implementing schemes to refine and optimize models using a gamut of drivers both macroeconomic and sentiments.

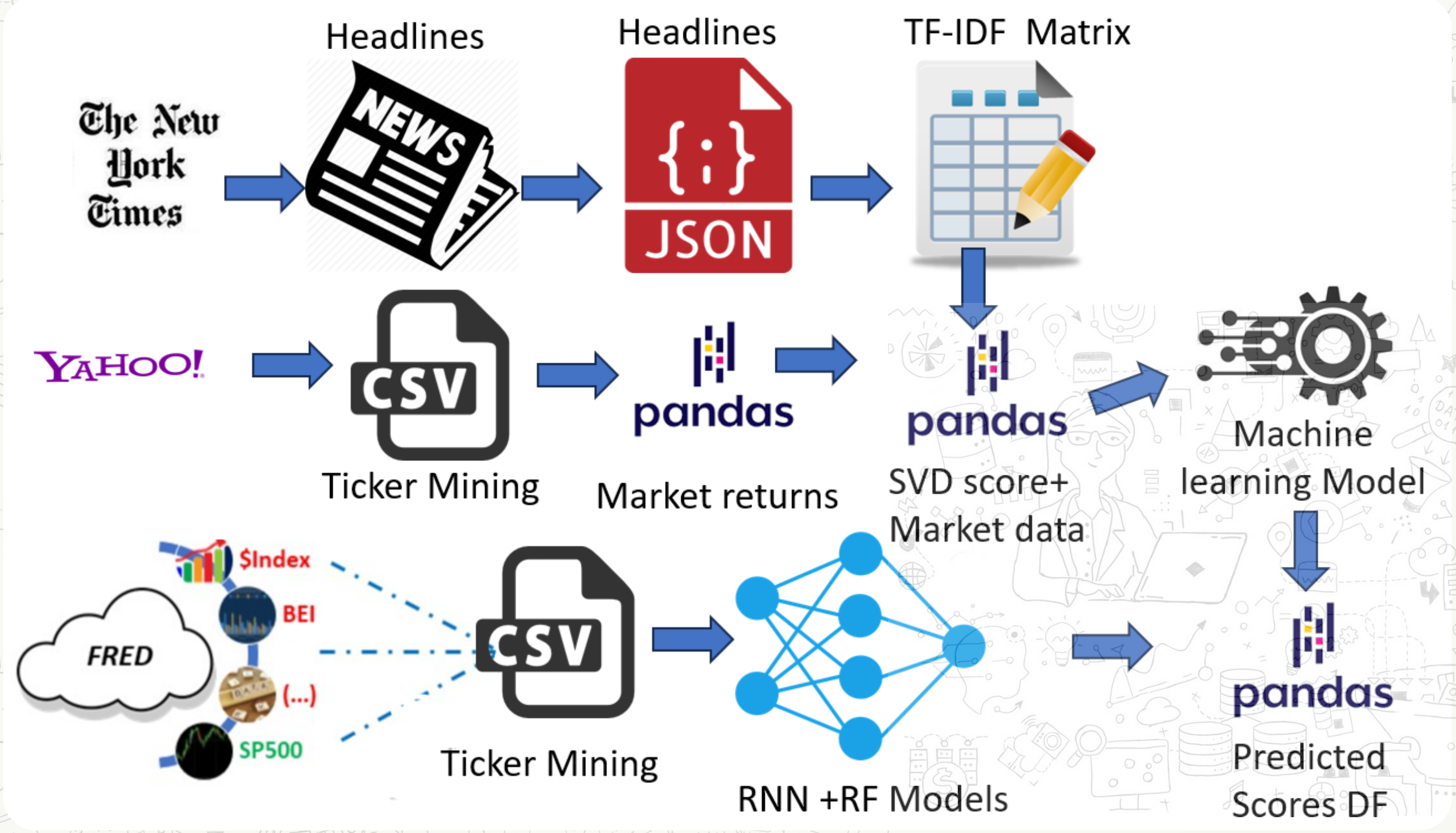
### Importance

- **Global Economic Impact:** S&P500's pervasive influence necessitates precise predictions.
- **Trillions Managed by Institutions:** Accuracy crucial for optimizing strategies in institutions handling trillions.
- **Profit Optimization:** Improved accuracy directly links to higher profits for investors and pension funds.
- **Innovative Tools:** Project innovation, like incorporating new variables and Natural Language Processing, enhances tools for financial decision-makers.



**Actual Value of the S&P 500 and the predicted value by the Multiple Linear Regression Model**

Smith, T. A., & Hawkins, A. (2015). An economic regression model to predict market movements.



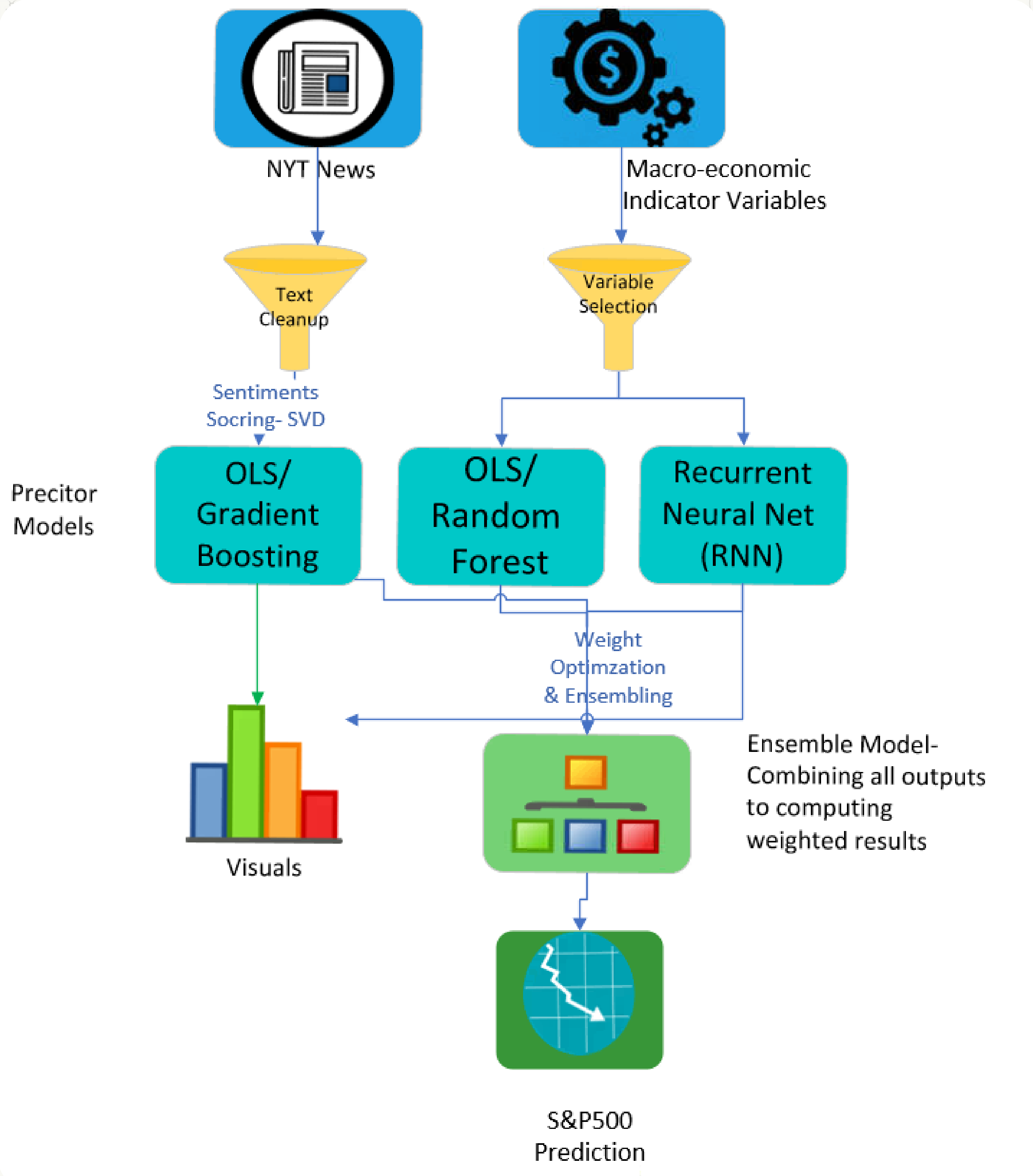
## Data

### DATA

- Socio-media input - New York Times News Headlines sentiment score via NYT API.
- Macro-economic data via FRED (D Prime Loan rate, financial stress, breakeven inflation, Oil and Gold ETFs for market volatility, and bond market indicators. etc ...) and Moody's APIs (PE ratios etc.).

### Data Characteristics and Wrangling:

- **News data → Text Format (JSON) → Vectorized using Term Frequency – IDF (Inverse Doc Hz).**
- **Vectorized News Data → Sentiment Scores using Singular Value Decomp (SVD)**
- **Macroeconomic Data → FRED API → CSV format**



## Approach

### Our Approach

- **Ensemble of the following models:**
  - GBM (Gradient Boosting Machine): Boosts accuracy by combining weak models. Handles complex relationships for better performance.
  - Random Forest: Combines diverse trees for robust predictions. Mitigates overfitting, captures patterns effectively.
  - RNN (Recurrent Neural Networks): Processes sequential data, capturing dependencies. Suitable for time series prediction, handles non-linear relationships.

### How does it work?

- GBM : Fits weak learners iteratively to correct errors. Combines diverse weak models for improved accuracy.
- Random Forest : Ensemble of decision trees using subsets. Aggregates results for a robust, high-performing model.
- RNN : Processes sequential data with hidden states. Maintains information from previous steps for time series prediction.

### Why our model is effective?

- The main intuition behind our approach is that a combination of multiple models capturing different relationships (both long and short term) can improve overall prediction and, in the course, have a more meaningful input to the decision-making process.

### What's new in the approach?

- Unlike traditional methods that rely solely on index values, our recurrent Neural Network (RNN) model incorporates macroeconomic variables, recognizing the impact of broader economic indicators on stock performance.
- Instead of conventional sentiment analysis tools, your model employs a unique approach using singular scores derived from Singular Value Decomposition of text news, adding a layer of sophistication to sentiment input.
- Addressing the limitations of traditional RNNs, this model integrates Long Short-Term Memory (LSTM) based RNN (Recurrent Neural Nets) to better capture long-range dependencies and improve sequential data modeling.
- Ensemble Model: The incorporation of different weights for averaging outputs from various models in the ensemble further enhances the predictive power of our innovative approach.

## Experiments and Results

### How did we evaluate our approaches?

- Models were evaluated on Mean Average (MAE), Root Mean Squared (RMSE), & Mean Squared (MSE) errors as accuracy metrices
- These are disparity measures between predicted & actual values, with lower error indicating a more robust model
- Best models amongst different algorithms (Random Forest, RNN, Linear Regression, and Gradient Boosting) .
- Evaluated the unbiased test error using Out of Bag (OOB) / Out of Time (OOT) datasets to address overfitting.

### What are the results? and How do our methods compare to other methods?

- **Experimental Setup:**
  - Two types of variables, macro-economic indicators, and sentiments from NYT news feed, were used to model S&P500 according to the above blueprint.
- Various modeling algorithms were applied to both sets of data, and model performance was assessed using error metrics for evaluating predictive accuracy to select best algorithm for the dataset
  - **Random Forest (Macro-Economic Data):** Outperformed Linear Regression in S&P500 prediction.
  - **RNN (Macro-Economic Data):** Captured complex non-linear relationships.
  - **News Sentiments (GBM):** Surpassed Linear Regression accuracy in daily stock return prediction.
  - **Ensemble Model:** Combined Random Forest, RNN, and GBM outputs with optimized weights. Reduced overall error rates drastically to 20%,enhancing predictive accuracy.
- **Comparison between Models and Ensemble Model:**
  - An ensemble model is created using weighted average for combining outputs. Final prediction involves solving a linear constrained optimization problem with the Simplex method. **Final Equation**: Predicted S&P500= a1*output_GBM + a2*output_RNN + a3*output_RF; **Optimized weights**: a1=0.015337; a2=0.472503; a3=0.51216
  - In Fig 1, blue solid line (ensemble prediction) is following the purple (actual) very closely, much closer than other model predictions, demonstrating the fact ensemble of models is a much strong predictor. In the adjoining table. We can see that error rates (CERRPCT for all forecast periods) for Ensemble model is remarkably low 20.2% which is much lower than the individual models ranging from 26% ~ 31%, making our ensemble model a unique concept and potent weapon for investors

Fig 1 Performance of individual & Ensemble Models

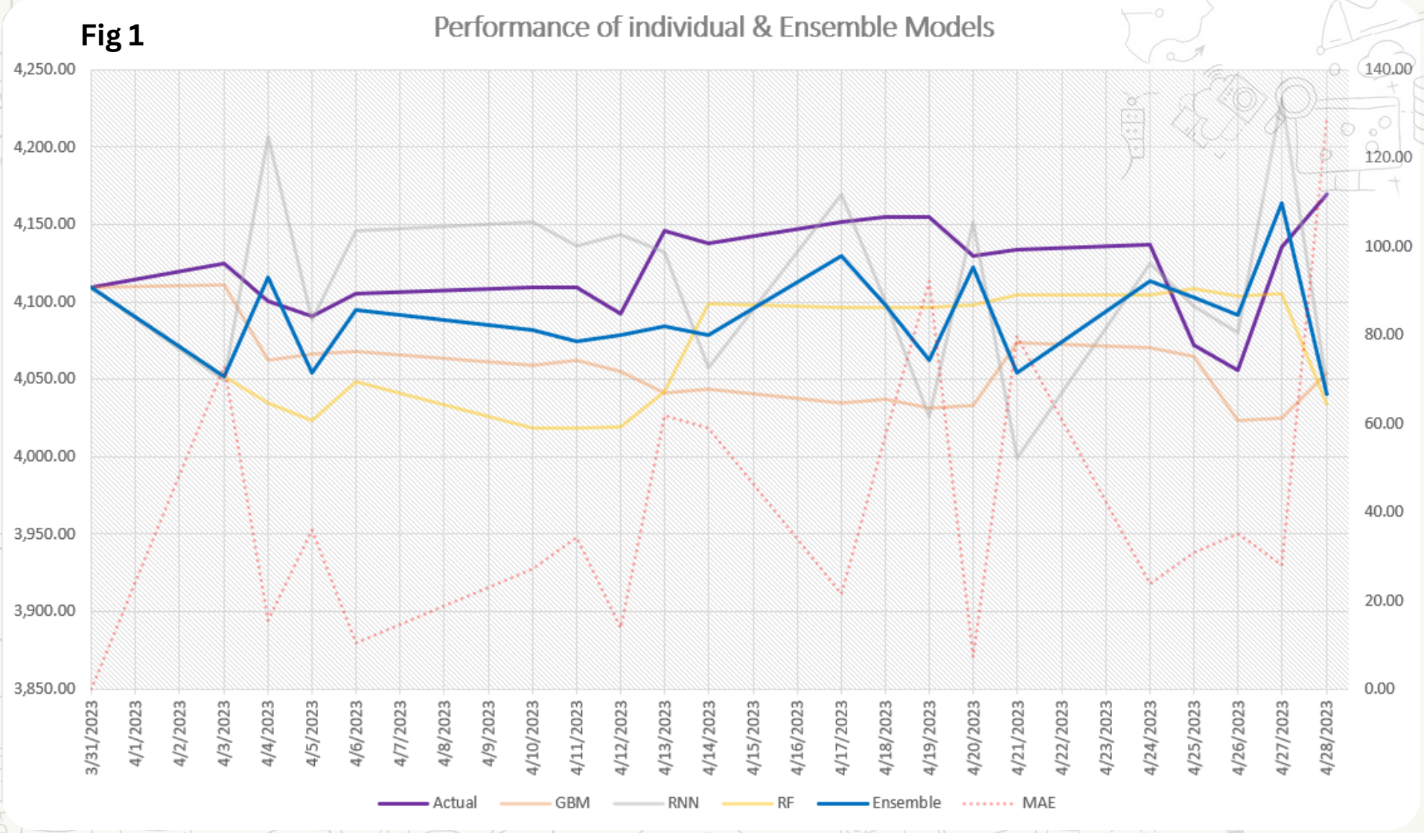
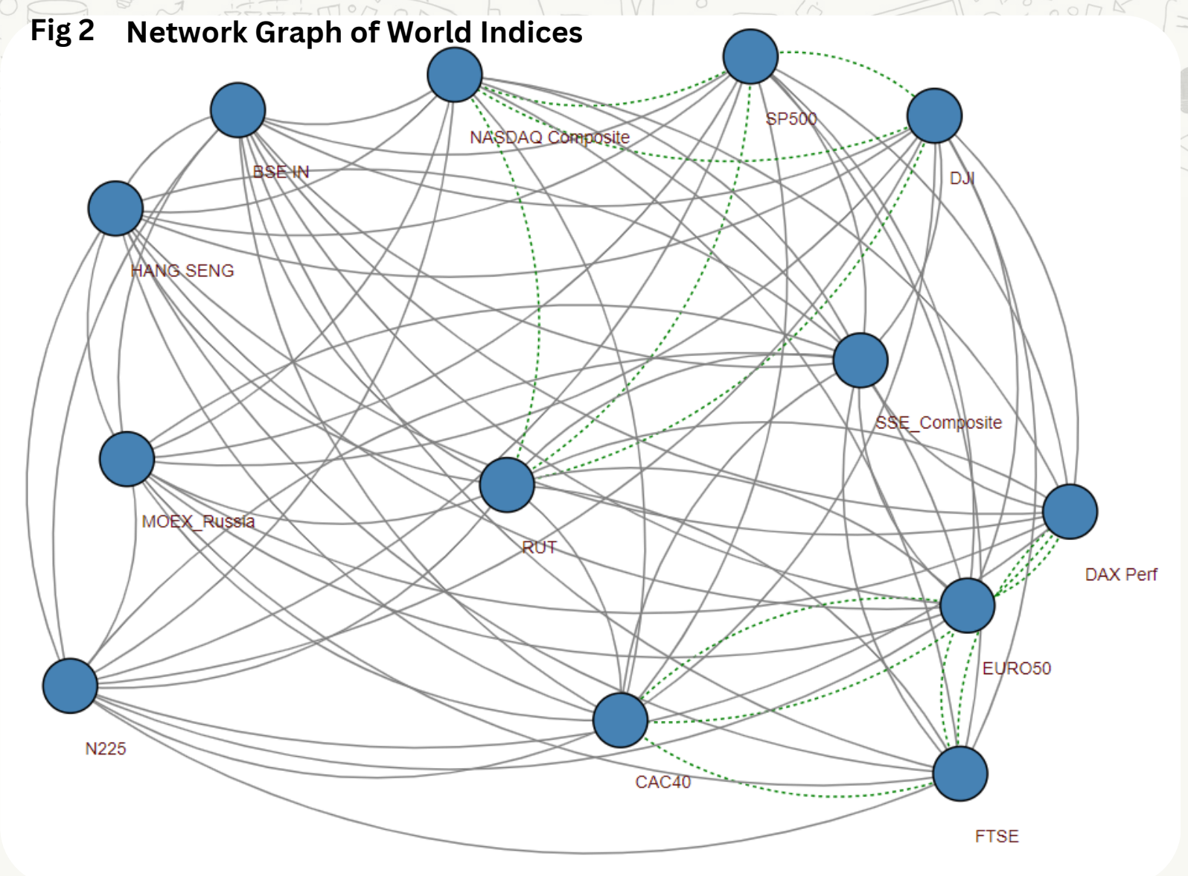


Fig 2 Network Graph of World Indices



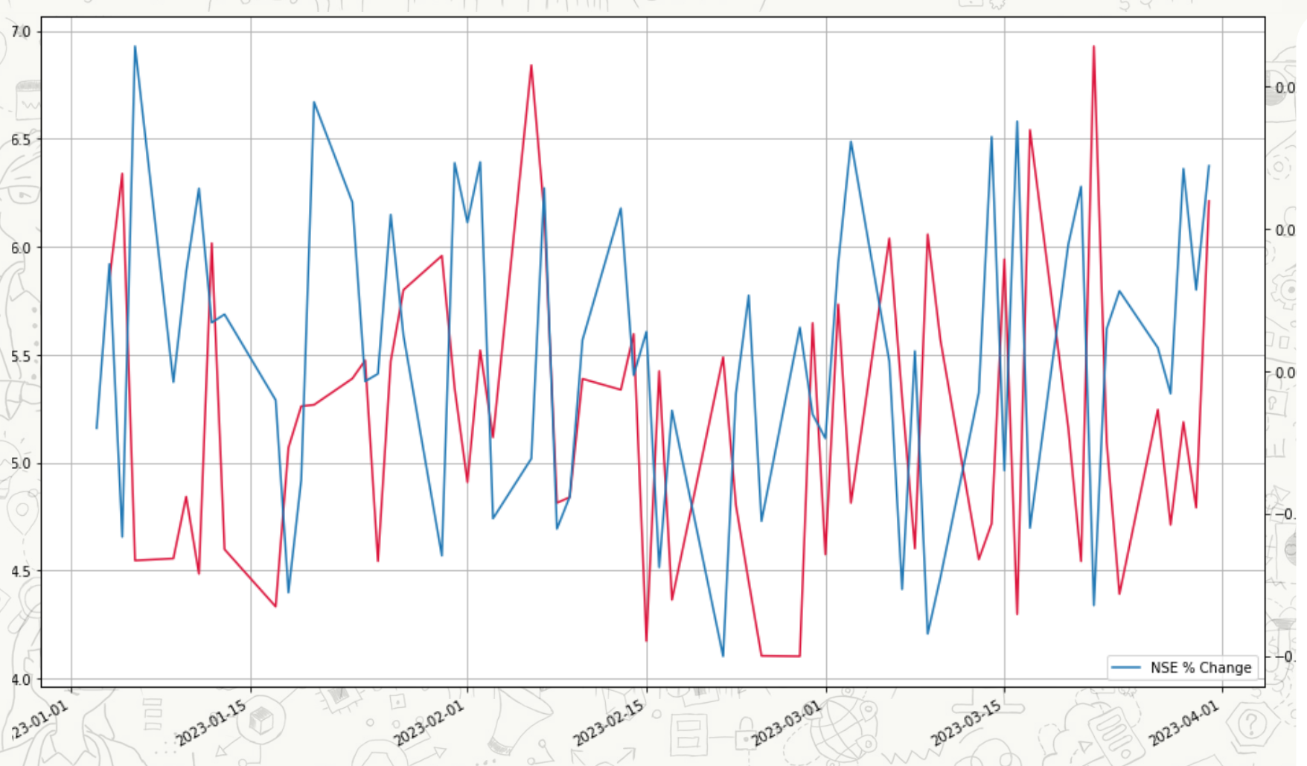


Fig 4 Bivariate Relationship Chart (Sentiments vs Daily % Change)

Fig 3 S&P500 Historic and Prediction

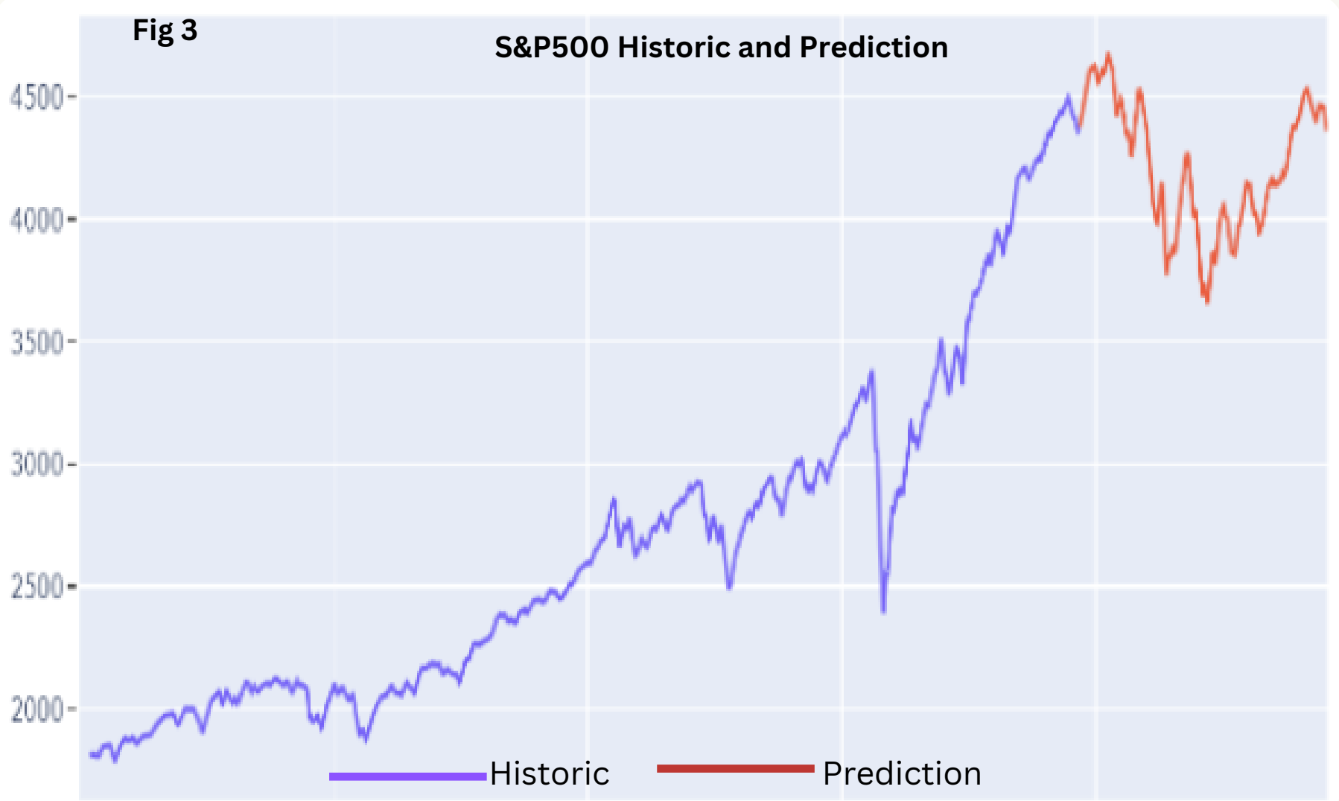


Table: Model Performance

*Ensemble Model Performance*

| CERRPCT_GBM | CERRPCT_RNN | CERRPCT_RF | CERRPCT_Ensemble |
|---|---|---|---|
| 31.2% | 26.1% | 28.5% | 20.2% |

*CERRPCT - Cumulative Error Percent Rate*