Linear Regression Project:
Predicting Stock Analyzing the Impact of Macroeconomic Indicators on SP 500 Performance

Lillian He, xh2652@columbia.edu
Youngshin Park, yp2691@columbia.edu
Sam Christianson, spc2157@columbia.edu
Amily Li, yl5711@columbia.edu

***Abstract*—This project investigates the interplay between S&P 500 performance and macroeconomic indicators to predict stock market trends and inform investment strategies. Utilizing datasets spanning 2011–2024, the analysis integrates descriptive statistics, categorical segmentation, and time-series visualizations to identify relationships between variables like GDP growth, unemployment, CPI, and effective federal funds rate. Linear regression was used to look for a relationship between macroeconomic indicatiors and S&P500 price. Due to the time series nature of this data set and a confounding effect of inflation, we decided simple linear regression would not be appropriate for this scenario. We tested several classification models including logistic regression, K-Nearest Neighbors, and Gradient Boosting to determine if we could predict a categorical increase or decrease in price the following month. Models were hyperparameters were tuned with cross-validation to determine optimal parameters and assess model accuracy. Results highlight challenges such as dataset imbalances and collinearity, but demonstrate moderate predictive capabilities. By leveraging these insights, the project aims to enhance understanding of market dynamics across varying economic conditions.**

I. Introduction

This project utilizes two key datasets, *S&P 500* and *Macro Data*, to explore the relationship between market performance and macroeconomic factors. The primarily objective is to analyze optimal investment timing, predict economic health for informed business decision-making, and capitalize on stock market volatility.

The S&P 500 Data encompasses comprehensive market performance metrics and investor sentiment analysis from 2011 to 2024, providing a robust framework for understanding stock market trends and behavioral finance. In parallel, the Macro Data provides insights into the broader economic environment, including indicators like inflation (CPI), economic growth(GDP), unemployment rates, and exchange rates.

In addition to descriptive and visual analyses, the project leverages categorical segmentation of economic variables, such as GDP growth and unemployment rates, to identify key patterns and trends across different economic conditions. For example, the data reveals distinct relationships between negative GDP growth, elevated unemployment rates, and stock market performance. Moreover, time-series visualizations of macroeconomic variables like CPI and EFFR provide valuable insights into the temporal dynamics of economic trends and their potential influence on market outcomes.

By integrating these two datasets with regression and classification models, the project aims to predict the directionality of the S&P 500 index in the subsequent month using macroeconomic data from the preceding month. By identifying the relationships, the project aspires to give stakeholders a better understanding of the market across various economic conditions.

II. Descriptive Analysis

A. Summary Statistics

The macroeconomic dataset contains 165 observations with several key quantitative variables. The Consumer Price Index (CPI) has a mean of 0.027 and a median of 0.021, indicating that the central tendency leans slightly below the mean due to a few higher values. The data distribution is slightly right-skewed, with an interquartile range (IQR) from 0.015 to 0.032. A maximum CPI of 0.091 suggests potential outliers. The Effective Federal Funds Rate (EFFR) exhibits a wide range, with a mean of 1.23 and a median of 0.18, reflecting a highly skewed distribution. The IQR, spanning from 0.09 to 1.88, and the maximum value of 5.33 highlight extreme variations over time. GDP Growth (%) centers near low growth values with a mean of 2.59% and a median of 2.8%, although extreme values, such as -28.1% and 35.2%, indicate significant outliers and fluctuations during severe economic events such as the Covid-19 pandemic. The unemployment rate ranges from 3.4% to 14.8%, with a mean of 5.55%, and most observations cluster between 3.8% and 6.8%, as shown by the IQR. Crude oil prices, measured through WTI (West Texas Intermediate), exhibit a relatively uniform distribution, with no significant outliers. The mean price is $71.51, with a maximum of $114.84. Both FX_EUR and FX_CNY exchange rates appear stable, with the Euro ranging between 0.675 and 1.02 and the Chinese Yuan between 6.05 and 7.31.

In the S&P 500 dataset, key stock market indicators include Price, Open, High, and Low values. These variables exhibit a strong right-skewed distribution, with a mean price of 2861.69 and a median of 2640.87. Prices have increased over time, reaching a maximum of 6034.91. The Change (%) variable reflects the market's volatility, with a mean of 1.01% and a range from -12.51% to 12.68%. This distribution centers

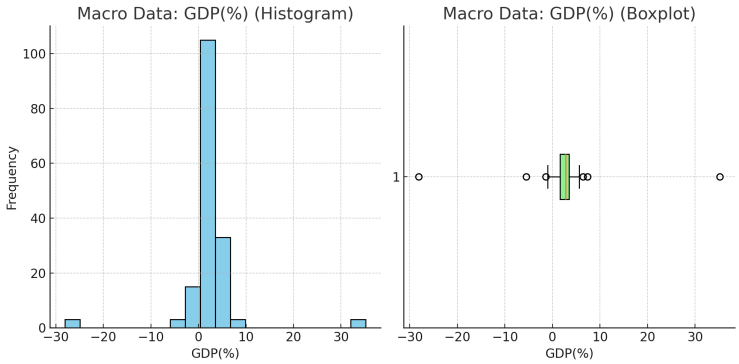
near the median of 1.59%, although extreme values suggest occasional high fluctuations.

B. Categorical Variables

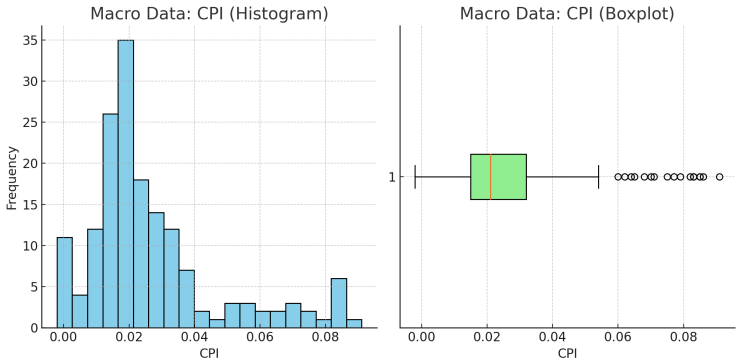
To better understand the relationships within the data, three categorical variables were created. GDP_Category divides GDP Growth into three bins: Negative, Low, and High. Most observations fall into the "Low" GDP category, indicating moderate economic performance over time. A smaller proportion belongs to the "Negative" and "High" categories, reflecting periods of economic downturn or exceptional growth. Unemployment_Category segments unemployment rates into Low (below 5%), Moderate (5-10%), and High (above 10%). Most observations fall in the "Low" and "Moderate" categories, with fewer cases in the "High" range. The Increase variable identifies whether the market's Change (%) was positive, with positive market changes dominating, aligning with the overall upward trend in stock prices.

C. Graphical Analysis

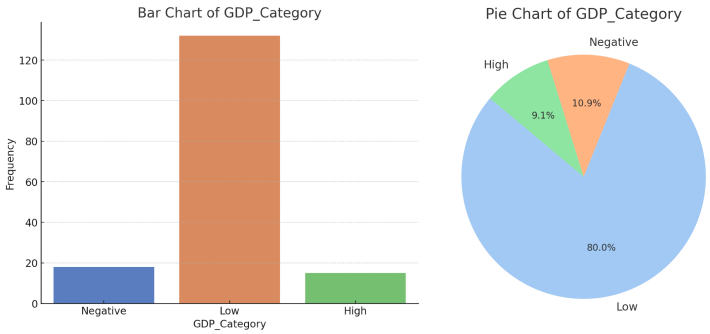
The quantitative variables in the dataset are well-represented through histograms and boxplots. For example, a histogram of GDP Growth reveals a right-skewed distribution with a concentration of values between 1.6% and 3.5%. Outliers, such as -28.1% and 35.2%, are particularly noticeable in the corresponding boxplot, which provides additional clarity on the spread of the data.



Similarly, the unemployment rate histogram shows a distribution centered around 5-6%, with higher outliers near 14.8% evident in the boxplot. A histogram of CPI illustrates inflationary pressures, while its boxplot highlights occasional extreme values. These visualizations are critical in understanding the variability and potential anomalies in the data.

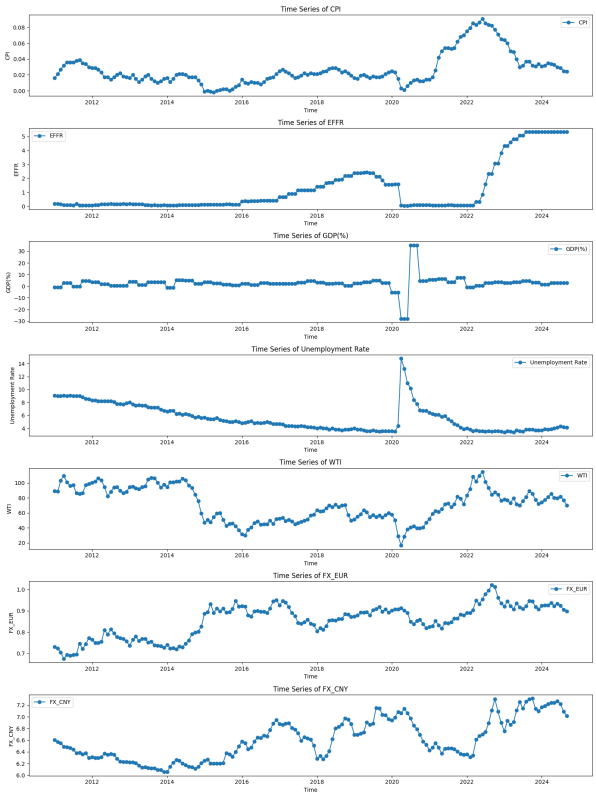


Bar charts and pie charts provide an overview of the categorical variables. For GDP_Category, a bar chart indicates that approximately 60% of the data falls into the "Low" category, while pie charts visualize that only 15% belongs to the "Negative" range. These proportions highlight the dataset's emphasis on moderate economic performance.



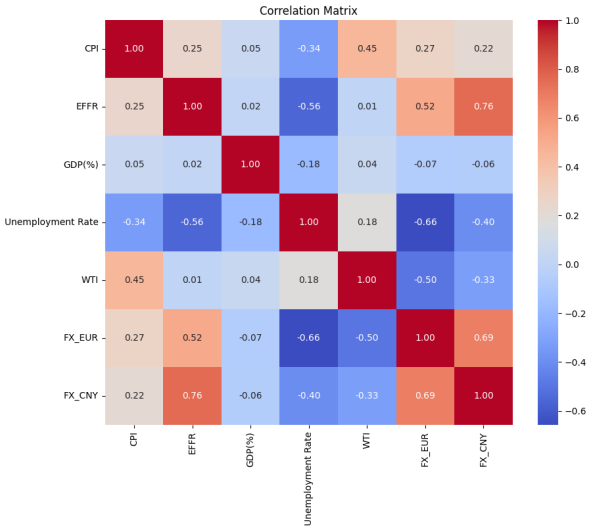
Unemployment_Category's bar chart confirms that more than 70% of the observations are in the "Low" or "Moderate" categories, indicating overall economic stability. The Increase variable's bar chart displays a significantly higher frequency of positive market changes, which is supported by the pie chart that visually emphasizes the dominance of the positive category.

Time series plots further contextualize the variability and trends over time for key macroeconomic indicators. For example, the CPI time series demonstrates gradual increases interspersed with occasional sharp spikes, highlighting inflationary trends during specific periods. Similarly, the EFFR time series illustrates extended periods of stability punctuated by abrupt policy-driven shifts. GDP Growth and unemployment rates exhibit complementary trends, where economic downturns are marked by dips in GDP Growth and corresponding peaks in unemployment. These time series visuals provide a longitudinal perspective, crucial for identifying patterns and cyclical behaviors in the data.



D. Correlation Analysis

The correlation matrix reveals significant relationships among key economic variables. FX_EUR, FX_CNY, and EFFR are highly correlated, reflecting interdependencies between interest rates and currency valuations. While such collinearity can complicate regression analysis by inflating standard deviations and obscuring the interpretation of coefficients, these concerns are mitigated in this study due to its predictive focus. Consequently, the emphasis remains on leveraging these interdependencies for accurate forecasts.



E. Descriptive Insights

The CPI and EFFR variables show moderate stability over time but can exhibit sudden spikes during significant economic events. High variability in EFFR suggests periods of aggressive monetary policy shifts. GDP Growth and unemployment rates exhibit strong relationships, with economic downturns marked by negative GDP growth and higher unemployment rates. The categorical breakdowns align with macroeconomic trends, providing a clear segmentation of economic conditions. The positive bias in the Increase variable aligns with the overall upward trajectory of stock prices over the observed period. Market volatility, as evidenced by extreme values in Change %, suggests sensitivity to external economic shocks.

F. Relationships for Further Analysis

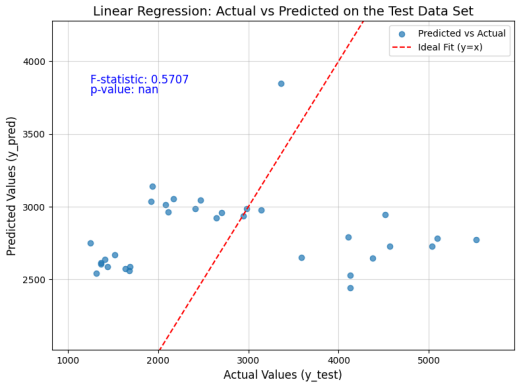
The analysis identifies several relationships for further exploration. Higher GDP growth is hypothesized to correlate with increased stock market prices or positive changes in performance. Similarly, unemployment rate categories may influence stock market volatility. The correlation matrix highlights strong associations between FX_EUR, FX_CNY, and EFFR, which could lead to unreliable coefficient estimates, inflated standard errors, and potential predictive inaccuracy if the collinearity changes. However, these concerns are not critical for our project since we prioritize prediction over inference, and we believe the collinearity is unlikely to change in future data. Lastly, examining CPI and EFFR alongside SP 500 trends could elucidate how macroeconomic variables drive market dynamics across diverse economic scenarios.

III. Statistical Analysis

A. Linear Regression

Linear Regression describes the relationship between an independent variable and a dependent variable. It is aim to predict the dependent variable based on the independent variable.

1) Cross-Validation Results: A linear regression model is implemented to evaluate the relationship between Price and predictors (CPI, EFFR, GDP(%), Unemployment Rate, WTI, FX_EUR, FX_CNY) and demonstrates strong explanatory power with an R-squared of 0.7956 and an adjusted R-squared of 0.7383. The statistical significance of the model is confirmed by an F statistic of 7.8407 and a p-value of 0.0000, allowing the rejection of the null hypothesis. Although the MSE of the test (333,643.20) is reasonable, the higher cross-validated MSE (2,197,220.65) suggests potential overfitting or model instability, which warrants further diagnostics. Overall, the model effectively captures the relationship between the features and the target, although enhancements such as feature engineering, residual analysis, or regularization could improve performance and robustness.



The linear regression model using PCA-transformed data (2 components) shows poor performance, with a Mean Squared Error (MSE) of 1,660,089.48 and a negative R-squared (-0.0171), indicating that the model explains less variance than a simple mean prediction. The Adjusted R-squared (-0.0849) further penalizes the inclusion of the principal components, confirming the lack of explanatory power. Cross-validation results are even worse, with a Cross-validated MSE of 3,299,169.08 and an extremely negative Cross-validated R-squared (-74.8586), highlighting instability and poor generalization. The low F-statistic (0.5707) and the absence of a valid p-value suggest that the model is not statistically significant. Overall, the PCA transformation appears to have oversimplified the feature set, leading to a failure to capture any meaningful relationship between the predictors and the target variable.

2) Model Evaluation:

Metric	First Model (PCA)	Second Model (No PCA)
MSE	1,660,089.48	333,643.20
R-squared	-0.0171	0.7956
Adjusted R-squared	-0.0849	0.7383
Cross-validated MSE	3,299,169.08	2,197,220.65
F-statistic	0.5707	7.8407
p-value	nan	0.0000

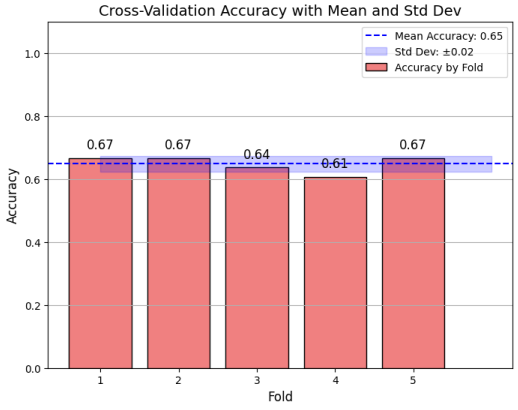
The model without PCA performs significantly better, with

higher explanatory power (R-squared = 0.7956), lower errors, and statistical significance; while the model with PCA simplifies the data but at the cost of accuracy and relevance, as evidenced by the negative R-squared and poor overall metrics. These two models have different results because PCA loses a large amount of variability in the original predictors by reducing the features to 2 components, which leads to poor model performance. The model without PCA retains all 7 predictors, which collectively have a strong relationship with the target variable, resulting in much better performance.

B. Logistic Regression

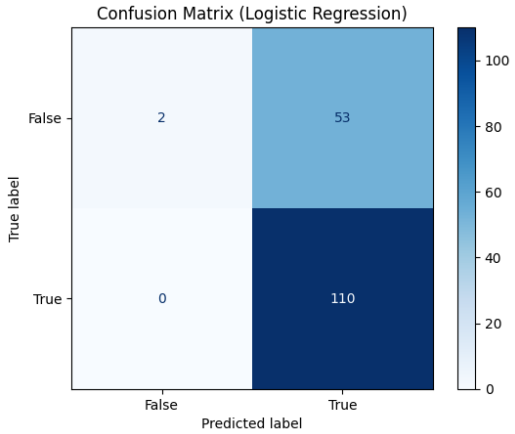
Noting that auto collinearity and potential confounding macroeconomic indicators, namely inflation, may impact the predictive power of this model over a scale as large as the one we are analyzing. Here we attempt to predict the month to month change of price as a categorical increase or decrease. We will use several different classification models. We will score each model based on accuracy and compare it to a naive model where we assume every month will increase. Based on our data set, our naive model has an accuracy of 66.7%. Logistic regression models the relationship between predictors and categorical response variables. Logistic regression estimates the probability that an observation falls into one of many categories of dichotomous variables, given continuous or categorical independent variables.

1) Cross-Validation Results: The logistic regression model was evaluated using 5-fold cross-validation, where the dataset was split into 5 parts. The model was trained on four parts and tested on the remaining part, repeating the process five times. The Cross-Validation Scores are: [0.6667, 0.6667, 0.6364, 0.6061, 0.6667]. This value indicates that the model correctly classified approximately 60.6% to 66.7% of the test instances.

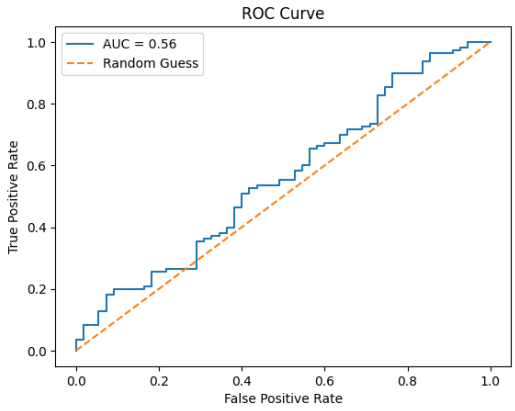


The average accuracy across all folds is 0.6485, which reflects the overall performance of the model in predicting whether the S&P 500 index goes "up" or "down" based on the macroeconomic conditions. Since the accuracy is not very high, there is some predictive power in the macroeconomic data, but may not be very strong.

2) Model Evaluation: The following confusion matrix shows the performance of the logistic regression model on the dataset. The model has a high recall for the "True" class, as there are no false negatives (FN = 0). However, it struggles with the "False" class, as most "False" instances (53 out of 55) are misclassified as "True". The matrix indicates a bias towards the "True" class, likely due to an imbalanced dataset or a threshold that favors "True" predictions.



The ROC (Receiver Operating Characteristic) curve evaluates the model's ability to discriminate between the two classes across different thresholds.



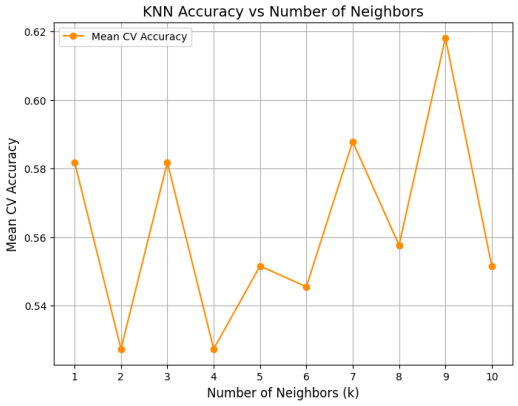
The AUC (Area Under the Curve) is 0.56, which is slightly better than random guessing (0.5) but indicates that the model has poor discriminatory power. Since the ROC curve is close to the diagonal, confirming that the model's predictions are only marginally better than random guesses.

C. K-Nearest Neighbors (KNN) Model

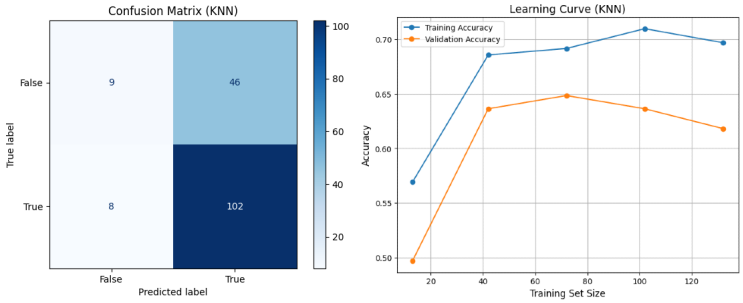
KNN is a machine learning algorithm that uses proximity to classify or predict the value of new data points. This model compares a new data point to a set of known data points to make prediction.

1) Cross-Validation Results: A K value of 9 is relatively moderate. It suggests that the model benefits from considering a balance between individual data points (lower K values) and broader trends in the dataset (higher K values). The best cross-validation accuracy is 0.6182, which suggests that the model correctly predicts about 62% of the instances during cross-validation.

2) Model Evaluation: The following graph shows how the mean cross-validation (CV) accuracy varies with the number of neighbors (k).



The optimal number of neighbors (k=9) corresponds to the highest mean CV accuracy of approximately 0.618 (61.8%). Small values of k (e.g., k=1, k=2) result in lower accuracy, possibly due to overfitting; larger values of k also result in reduced accuracy, likely due to underfitting as the model becomes too generalized.



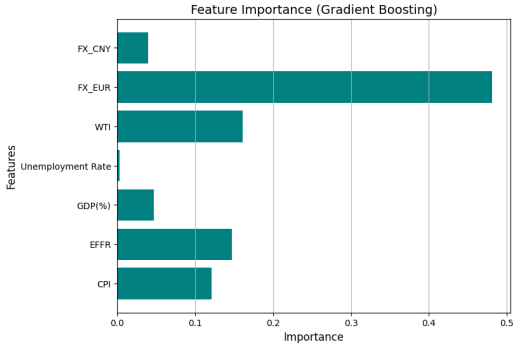
According to the Confusion Matrix (KNN), the high number of False Positives (46) suggests the model tends to overpredict the "True" class, which may indicate an imbalance in the dataset or a need for improved feature selection.

The training accuracy is consistently higher than the validation accuracy, peaking at around 0.72. This indicates the model performs well on the training set. The validation accuracy peaks around 0.65 with a medium-sized training set, then decreases slightly as the training size increases. The consistent gap between training and validation accuracy suggests a degree of overfitting; the model generalizes less effectively to unseen data.

D. Gradient Boosting

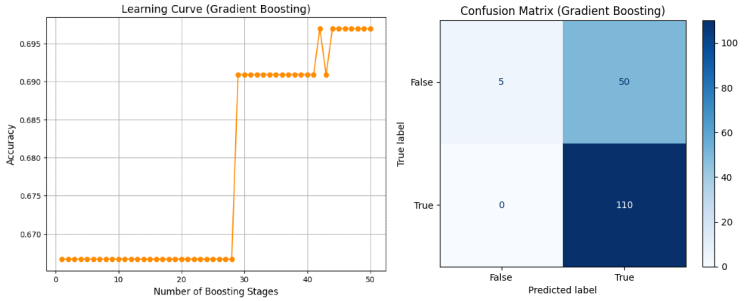
Gradient Boosting combines multiple weak models to create a single, more accurate predictive model.

1) Cross-Validation Results: The model achieved an overall accuracy of 69.7%, which means that about 70% of the predictions were correct.



The plot highlights the features which are most influential in the Gradient Boosting model. The top features are FX_EUR and WIT(Crude Oil Price).

2) Model Evaluation: The following plots shows the Gradient Boosting evaluation results.



The left plot shows the model's accuracy on the test data as a function of the number of boosting stages (iterations). At the initial stage, accuracy remains stagnant at around 67%, indicating that the model isn't improving during the initial iterations. Around stage 30, there is a significant jump in accuracy, suggesting the model starts to capture more meaningful patterns. Accuracy stabilizes close to 69.5% by boosting stage 40, with slight fluctuations.

The confusion matrix on the right evaluates the model's predictions on the test set. The model perfectly identifies all "True" instances (False Negatives = 0), achieving a recall of 100% for the "True" class. However, it struggles with the "False" class, misclassifying 50 out of 55 instances as "True" (very high false positives), resulting in poor precision and recall for the "False" class.

E. Final Results and Comparison

IV. Discussion and Conclusion

This project explored the relationships between macroeconomic indicators and S&P 500 performance using various statistical and machine learning techniques. The analysis underscored the importance of understanding interdependencies among economic variables to predict stock market trends effectively.

The descriptive and graphical analyses highlighted key patterns in the data, such as the cyclical nature of GDP growth and unemployment rates and their influence on market dynamics. Notably, the correlation analysis revealed strong associations between interest rates, currency fluctuations, and stock market performance, emphasizing the interconnected nature of these variables.

Among the models tested, linear regression on first glance demonstrates strong explanatory power, with an R-squared of 0.7956, showing a degree of accuracy in predicting relationships between predictors and the target variable. However, we note that there are effects relevant to this data that simple linear regression does not account for, including auto-collinearity in the data and inflation. The model also exhibited potential overfitting, as evidenced by the high cross-validated mean squared error (MSE). Instead, classification models were tested to determine whether we could predict the sign of the change in price from month to month using several of the previous

month's macroeconomic indicators. Logistic regression provided limited predictive power, with an average accuracy of 64.85%, while the K-Nearest Neighbors model achieved a peak validation accuracy of 61.8%. Gradient Boosting emerged as the most effective predictive model, achieving an overall accuracy of 69.7% and highlighting the importance of iterative refinement in capturing meaningful patterns. Gradient boosting was the only model to beat the Naive model of assuming all months would increase in price which has an accuracy of 66.7% based on this data set.

Despite these successes, the project faced challenges such as dataset imbalances and multicollinearity, which complicated model interpretation and reduced generalizability. These limitations underscore the need for more robust preprocessing techniques, such as resampling strategies to address imbalances and regularization methods to mitigate collinearity.

Future research could explore incorporating additional macroeconomic variables, refining feature selection processes, and testing more advanced ensemble methods. By addressing these aspects, predictive models could achieve greater accuracy and reliability, further supporting stakeholders in making informed investment decisions.

In conclusion, this project provided valuable insights into the interplay between macroeconomic factors and stock market performance. While the models demonstrated moderate predictive capabilities, the findings lay a foundation for further exploration of market dynamics under varying economic conditions.

APPENDIX

A. Datasets

- S&P 500 (SPX) historical data.
- Consumer Price Index(CPI) historical data .
- Effective Federal Funds Rate (EFFR) historical data .
- GDP historical data .
- Unemployment Rate historical data .
- West Texas Intermediate (WTI) historical data .
- Euro Exchange Rate historical data .
- Chinese Yuan Exchange Rate historical data .

B. Implementation

This Colab notebook contains the code used to implement our models.