Forecasting S&P 500 using macroeconomic data

Songyan Xu (sx274) Purvesh Jain (pj268)

Course Code: 5741
Github Link : https://github.com/leoleoeo666/5741-Project.git
Video Presentation: ORIE 5741.mp4

Introduction

This report is dedicated to a detailed analysis and forecasting of US financial market trends through a blend of statistical analysis and machine learning methods. The core objective of this project is to delve into the relationships among various economic indicators to develop predictive models. These models are designed to provide insights into potential future market behaviors, assisting in more informed decision-making processes for investors, analysts, and policy makers.

Dataset Description

The dataset utilized in this report comprises six key variables: S&P 500 Prices, Gold Prices, USD Index, WTI Prices (Crude Oil), 3-Month Treasury Rate, and 10-Year Treasury Rate. This data has been gathered from two prominent sources: Federal Reserve Economic Data (FRED) and Yahoo Finance. The selection of these sources ensures the reliability and accuracy of the data, which spans from April 2014 to April 2024. We choose to use daily frequency and only uses data that are published daily, which provides a granular view of market dynamics.

Data Cleaning and Exploratory Data Analysis

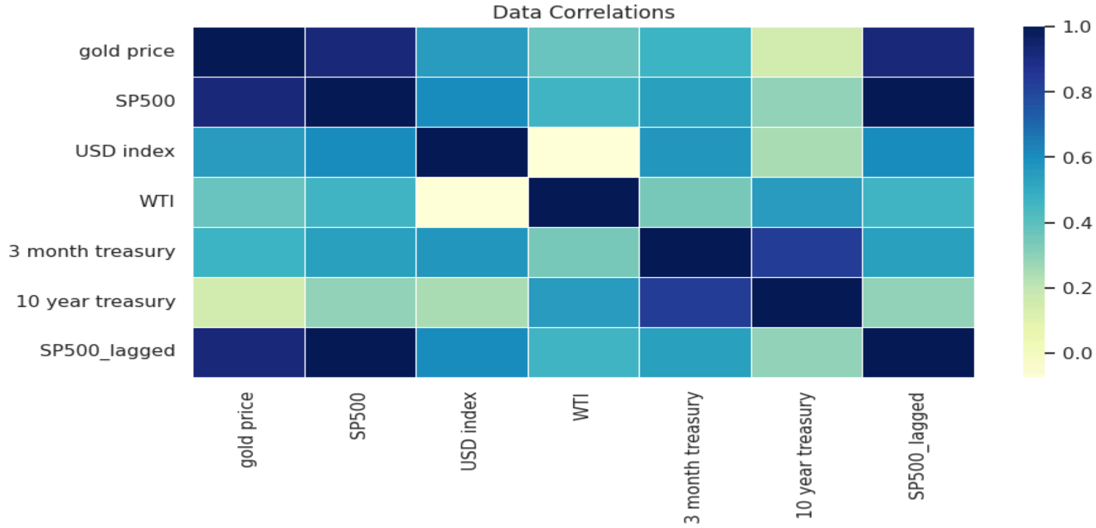
Due to different market closing times and diversity of our data, we do have some missing data which we handled using a forward filling method. We also create another column S&P lagged, ensuring our models are predicting future market movement. Below is a brief look at our data after dropping NA values. The train-validation-test data split we use is 60-20-20.We also standardize features by removing the mean and scaling to unit variance using the StandardScaler() module in python.

Date	Gold	SP500	USD_Index	WTI	3_month_treasury	10_year_treasury	SP500_lagged
2014-04-14	1326.40	1862.31	93.8046	104.05	0.04	2.65	1862.31
2014-04-15	1302.65	1862.31	93.9486	103.70	0.04	2.64	1862.31
2014-04-16	1302.80	1862.31	93.9461	103.71	0.04	2.65	1864.85
2014-04-17	1294.85	1864.85	93.8936	104.33	0.03	2.73	1871.89
2014-04-18	1293.78	1871.89	93.9629	104.35	0.04	2.73	1871.89

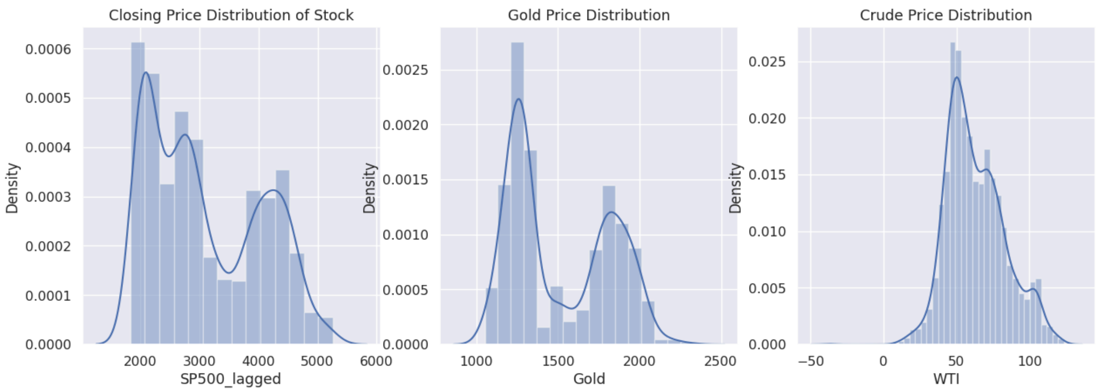
Relationship between factors and S&P 500

We then plot our independent variables, 'Gold', 'USD_Index', 'WTI', '3_month_treasury', and '10_year_treasury' against 'SP500_lagged'. Based on the graphs, all our variables, especially the gold price, tend to have a positive relationship with 'SP500_lagged'.

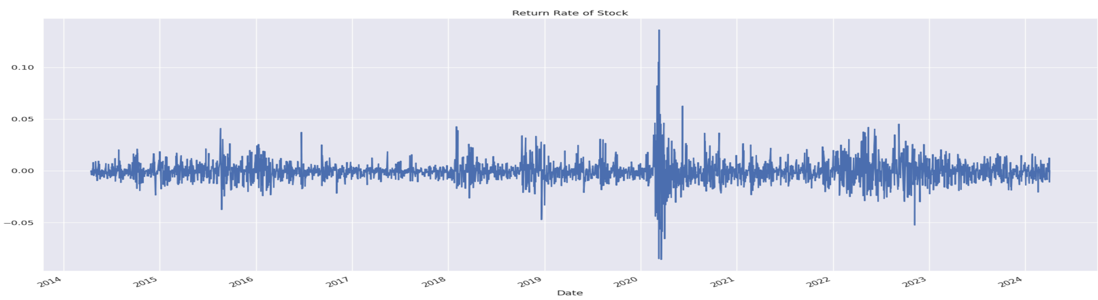
Correlations between features



Price Distribution



Daily Returns



Feature Engineering

S&P 500 is by nature volatile, and we therefore create 20-, 50- and 100-days moving averages for SP500_lagged to smooth the price and cut down noise. Below is a brief look at our data after dropping NA values.

Date	Gold	SP500	USD_Index	WTI	3_month_treasury	10_year_treasury	SP500_lagged	MA_20_days	MA_50_days	MA_100_days
2014-08-29	1287.57	2003.37	94.7251	97.86	0.03	2.35	2002.28	1966.5155	1967.3644	1935.2807
2014-09-01	1287.30	2002.28	95.0630	92.92	0.03	2.42	2002.28	1970.6190	1968.4104	1936.6804
2014-09-02	1265.90	2002.28	95.0630	92.92	0.03	2.42	2000.72	1974.6430	1969.2342	1938.0645
2014-09-03	1269.07	2000.72	94.8709	95.50	0.03	2.41	1997.65	1979.0470	1970.0428	1939.3925
2014-09-04	1261.10	1997.65	95.2641	94.51	0.03	2.45	2007.71	1982.8530	1970.9778	1940.7507

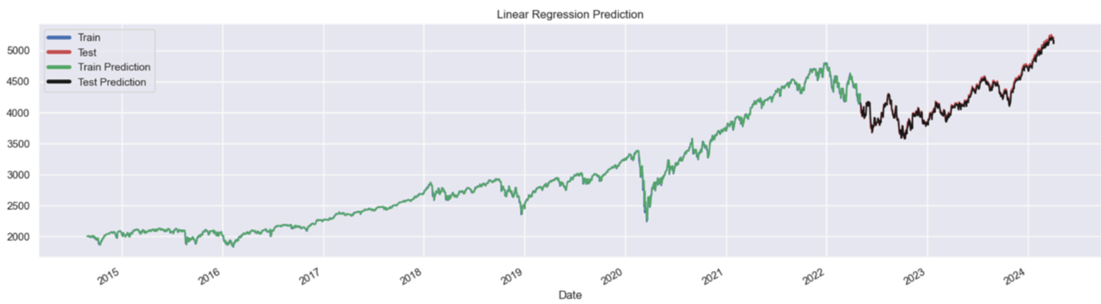
Machine Learning Models

We try out multiple techniques to find the best model to predict S&P 500 returns

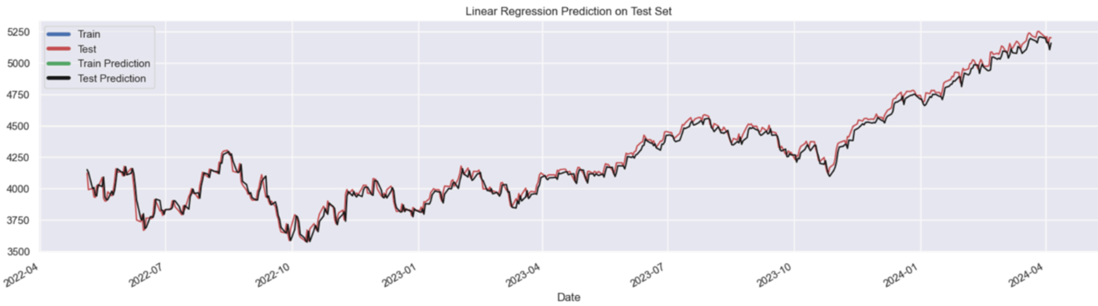
Linear Regression

Description	A statistical method for modeling the relationship between a dependent variable and one or more independent variables.
Pros	Effectiveness for predicting outcomes based on linear relationships; quick calculation times.
Cons	Can perform poorly if the assumptions of linear regression are not met.
Robustness	Generally low unless modifications like ridge or lasso are applied to handle multicollinearity and overfitting.
Suitable for	Data with clear, linear relationships between features and target.

RMSE Train	RMSE Test	R2 Train	R2 Test
32.57	50.99	0.99	0.98

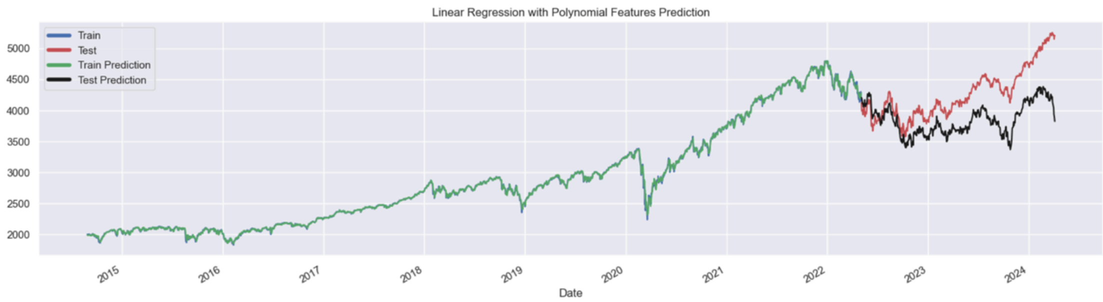


Model demonstrates high predictive accuracy and good generalization from training to testing, with slight indications of overfitting reflected in the increase in RMSE from training to testing.



Linear Regression with Polynomial Feature

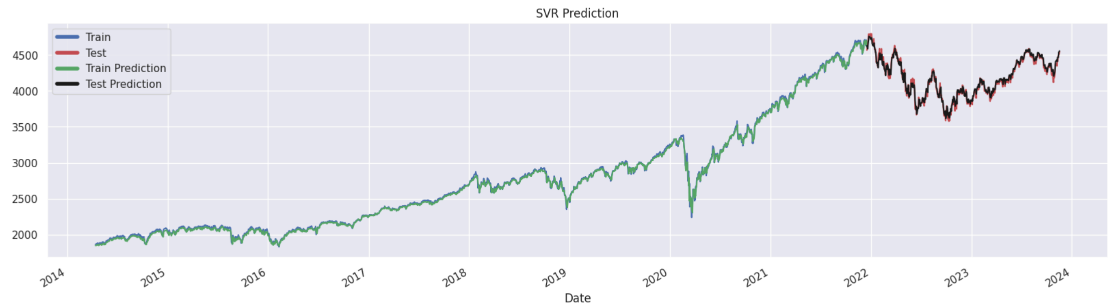
Adding another layer of complexity, when we incorporated polynomial features to delve into more intricate relationships, the model began to significantly overfit. The RMSE for the training data modestly decreased to 30.91, signaling a better fit on the training set. However, this improvement was misleading as the RMSE for the test data escalated dramatically to 500.03. Furthermore, the R-squared value for the test data became negative, dropping to -0.59, clearly demonstrating that the model with polynomial features failed to generalize effectively to new data."



RMSE Train	RMSE Test	R2 Train	R2 Test
30.91	500.03	0.99	-0.59

SVR

Description	SVR uses the principles of support vector machines for regression, by finding a hyperplane that best fits a given set of points with a margin of error.
Pros	Effective for both linear and non-linear data, offering flexibility with kernel choices to capture complex relationships.
Cons	Complex to implement & computationally expensive for large datasets due to the need to invert a matrix.
Robustness	High robustness against overfitting, particularly in scenarios where the number of dimensions exceeds the number of samples.
Suitable for	Data with clear, linear relationships between features and target.

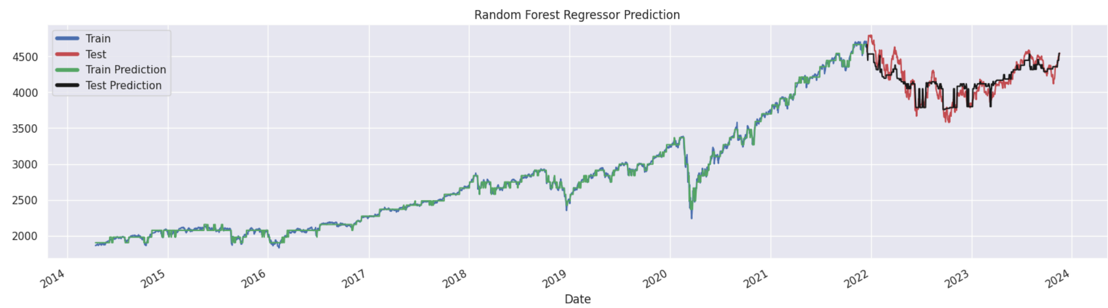


RMSE Train	RMSE Test	R2 Train	R2 Test
41.71	70.24	0.99	0.97

The SVR model exhibits strong predictive accuracy with robust training performance, though the increase in RMSE from training to testing suggests modest overfitting.

Random Forest

Description	Ensemble learning method that constructs multiple decision trees during training and outputs the mode of the classes or mean prediction of the trees.
Pros	Excellent for handling datasets with high dimensionality and can manage thousands of input variables without variable deletion.
Cons	Due to its ensemble nature, it can require significant computational resources and time to train, especially with very large data sets.
Robustness	Highly robust against overfitting compared to single decision trees, especially with more trees in the forest.
Suitable for	Performs well with large datasets with complex relationships that may involve interactions and non-linearities.

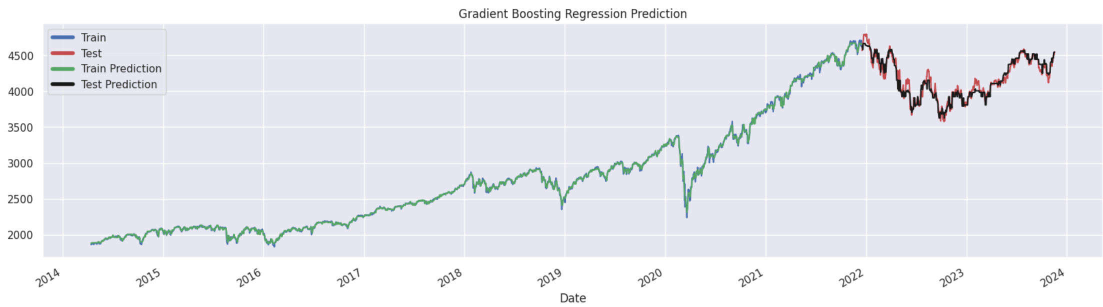


RMSE Train	RMSE Test	R2 Train	R2 Test
36.74	178.73	0.99	0.80

The Random Forest model shows excellent training performance but a significant increase in RMSE and drop in R^2 on the test data indicate considerable overfitting.

Gradient Boosting Regressor

Description	Builds an additive model in a forward stage-wise fashion; it allows for the optimization of arbitrary differentiable loss functions.
Pros	Excellent handling of non-linear relationships; robust to outliers in output space.
Cons	Can overfit on very noisy datasets and requires careful tuning of parameters.
Robustness	High robustness in general, particularly with sufficient data and proper tuning.
Suitable for	Works well with complex datasets that exhibit non-linear patterns.

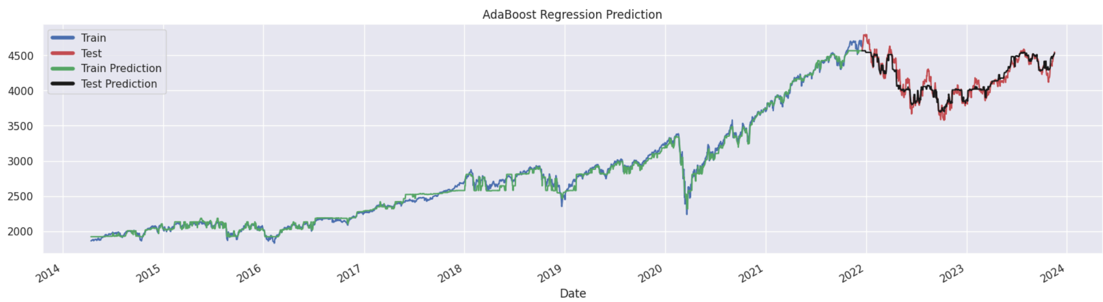


RMSE Train	RMSE Test	R2 Train	R2 Test
25.44	174.61	0.99	0.81

The Gradient Boosting Regressor achieves high training accuracy and maintains strong generalization to test data, with a modest increase in RMSE suggesting slight overfitting.

AdaBoost Regressor

Description	Ensemble technique that combines multiple weak learners to create a strong predictive model, adjusting weights of incorrectly predicted instances.
Pros	Can significantly improve the performance of weak learners; less prone to overfitting compared to other algorithms if weak learners are simple.
Cons	Potentially less effective on datasets with a high level of noise.
Robustness	Generally robust to overfitting in low-noise scenarios.
Suitable for	Effective for binary classification & regression with moderate dataset sizes, features

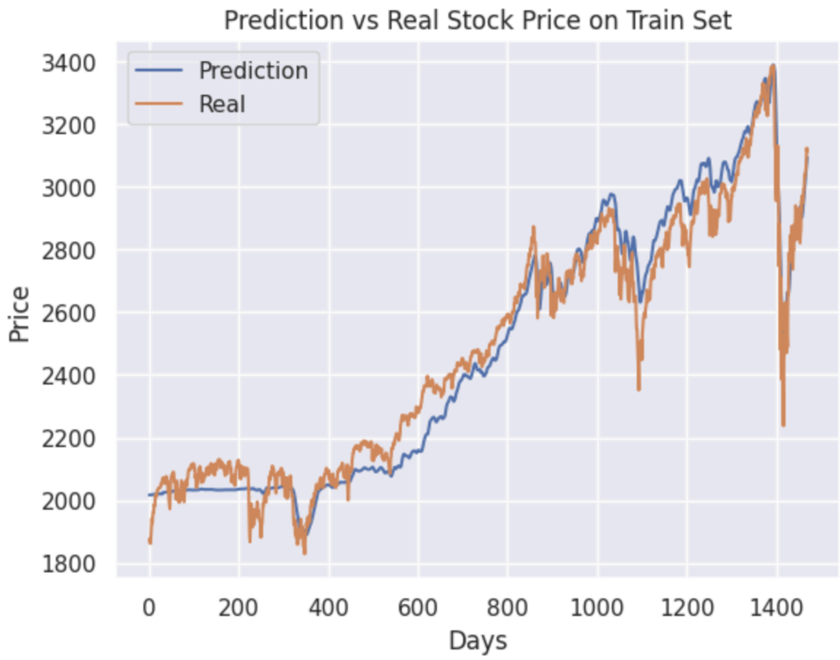


RMSE Train	RMSE Test	R2 Train	R2 Test
61.25	189.77	0.99	0.77

The AdaBoost Regressor shows excellent training performance, but a noticeable increase in RMSE on the test data indicates mild overfitting while maintaining a strong test R^2 score.

LSTM

Description	Recurrent neural network(RNN) designed to avoid long-term dependency problems, making it capable of learning order dependence in sequence prediction problems.
Pros	Highly effective for sequence prediction, time series analysis, and natural language processing due to its ability to maintain state over long sequences.
Cons	Complex model architecture leads to high computational cost and long training times.
Robustness	Highly robust against the vanishing gradient problem, allowing it to learn from long data sequences effectively.
Suitable for	Particularly well-suited for data where the context or sequence order is important, such as time series data, speech, and natural language.



RMSE Train	RMSE Test	R2 Train	R2 Test
78.24	553.11	0.96	0.96

The LSTM model maintains consistent R^2 scores from training to testing, suggesting stable performance, but the high RMSE values indicate a potential need for model tuning or reevaluation of the data used.

Conclusion

In conclusion, the content of this report demonstrates varying degrees of success among different models, highlighting the critical role of model selection and parameter tuning. While models like Linear Regression and LSTM showed potential of success, they also exhibited challenges like overfitting, which emphasize the need for rigorous validation and testing. Models such as Linear Model with polynomial feature, Gradient Boosting and Random Forest offered robustness with training data but with limitations in generalization to unseen data, as shown in the much higher RMSE in test data compared to training data. This analysis not only underscores the intricate relationships between economic indicators and market performance but also offers valuable insights for enhancing predictive accuracy in financial market forecasting, which is crucial for prudent investment and policy-making decisions.

Contribution

Songyan Xu: Data Gathering, Data Cleaning, EDA, Linear Regression,
Purvesh Jain: SVR, Random Forest, Gradient Boosting, AdaBoost, LSTM
All group members agree they work fairly and equally