Q4/December 2025

Economic Factors and the S&P 500

DATA 201 - Data Analytics & Machine Learning

Created by Ezey Duru and Erika Holbrook

Our *Team*

Ezey Duru

Data Scientist and Analyst

Ezenwanyi.Duru@tufts.edu

Erika Holbrook

Data Scientist and Analyst

Erika.Holbrook@tufts.edu

Presentation Contents

Background

The Great Recession

- Collapse of the U.S. housing market and subprime mortgage defaults
- Failure of mortgage-backed securities
- Credit markets froze major banks required bailouts
- S&P 500 fell over 50%,
- One of the worst financial crises since the Great Depression

Market conditions can shift, this is why analyzing economic indicators is essential for predicting trends

Data Cleaning

Economic Feature Indicators Summary Table

Corporate Profits	Total earnings of U.S. corporations.	Real Estate Index	Trends in property price changes.
Crude Oil Price	Market price of a barrel of crude oil.	Retail Sales	Total consumers spending.
Federal Interest Rate	Borrowing rate between banks set by the Federal Reserve.	Tech Investments	Capital spent on technology and innovation.
GDP (Gross Domestic Product)	Total value of goods and services produced in a country.	Unemployment Rate	Percentage of people without jobs
Gold Price	Market price of one ounce of gold.	Inflation Rate	Rate at which overall prices for goods and services rise.

yahoo! finance

FRED® ECONOMIC DATA | ST. LOUIS FED

Machine Learning

Data Cleaning & Preprocessing

Data Collection & Merge

- Loaded 11 FRED CSVs
- Renamed columns
- Merged features + S&P 500 (^GSPC, Yahoo Finance) on date

01

Missing Data & Filtering

- Converted dates
- Kept 2006–2010 window
- Forward-filled monthly, quarterly, yearly macroeconomic indicators
- Dropped rows with missing S&P data

02

Feature Engineering

- Computed 200-day moving average
- Created daily signal variable
- Shifted target signal for prediction

03

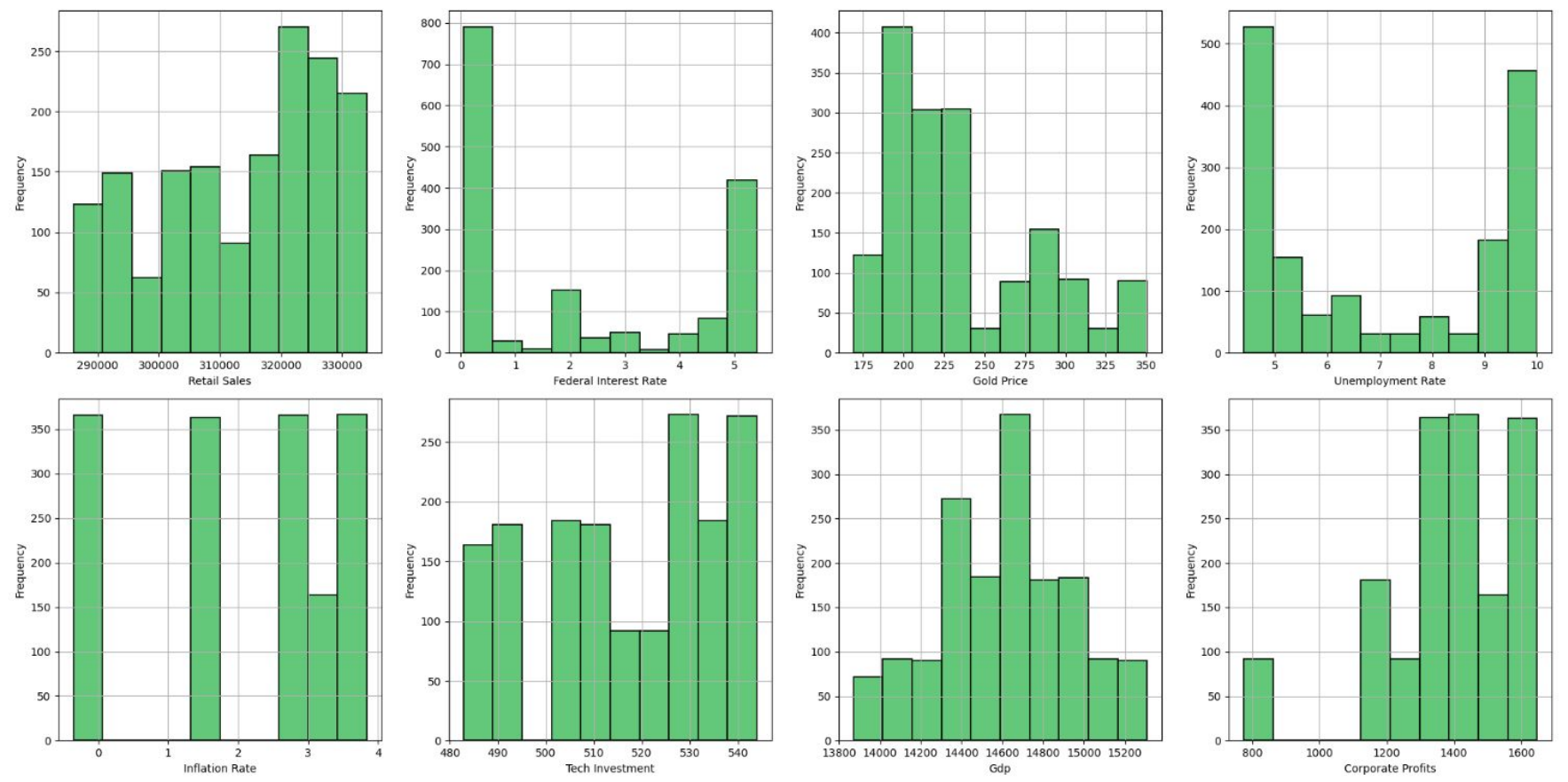
Final Dataset Preparation

- Saved cleaned dataset
- Verified no nulls and correct types
- Used for model training

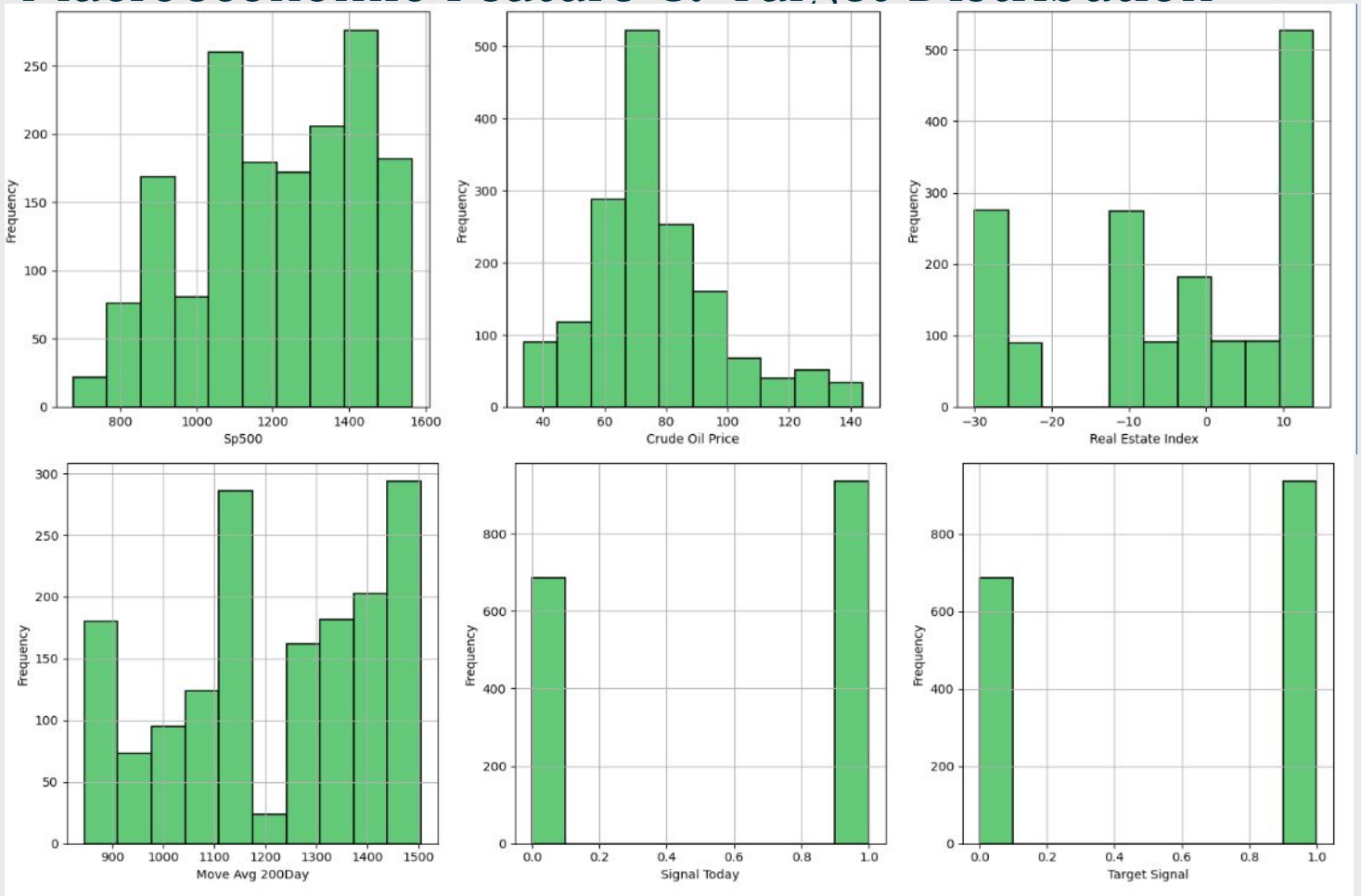
04

Data Visualization

Macroeconomic Feature Distribution

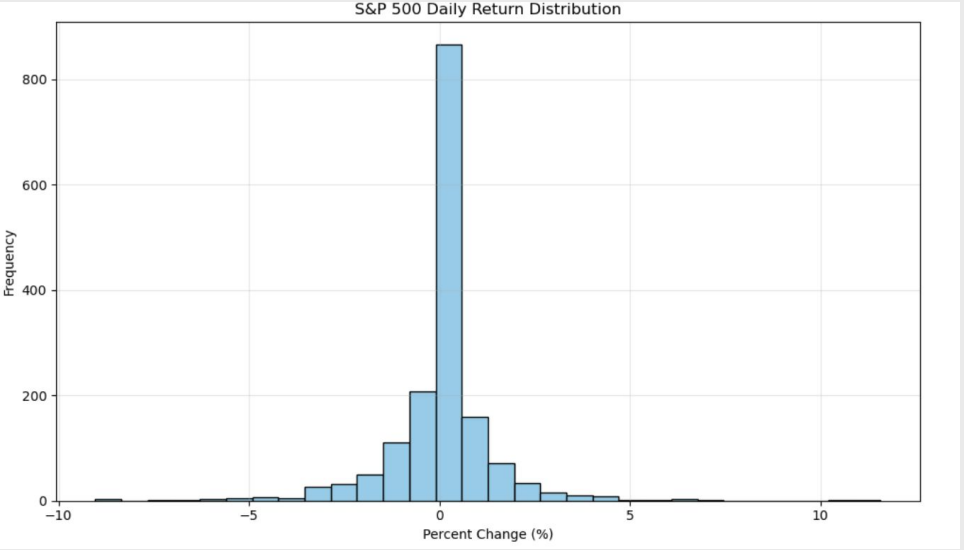


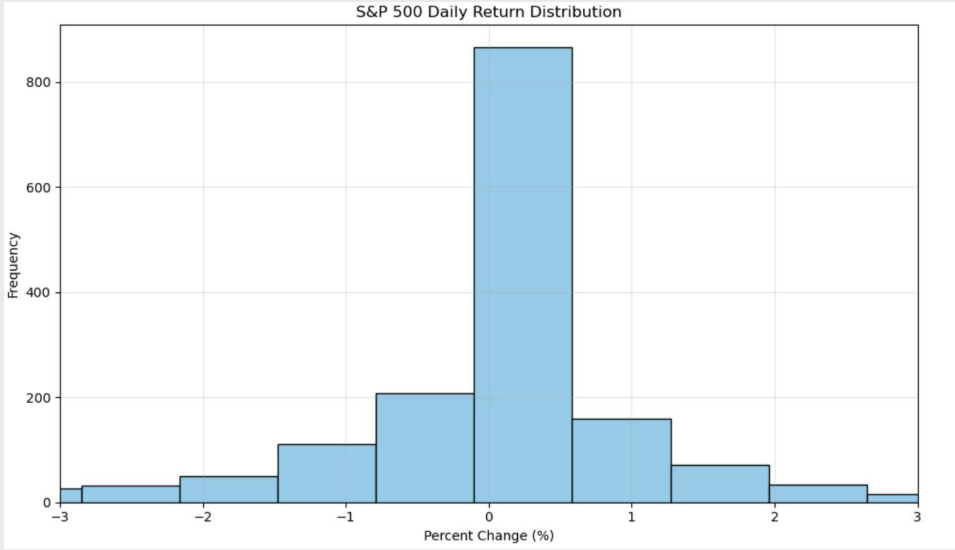
Macroeconomic Feature & Target Distribution



Target Prediction

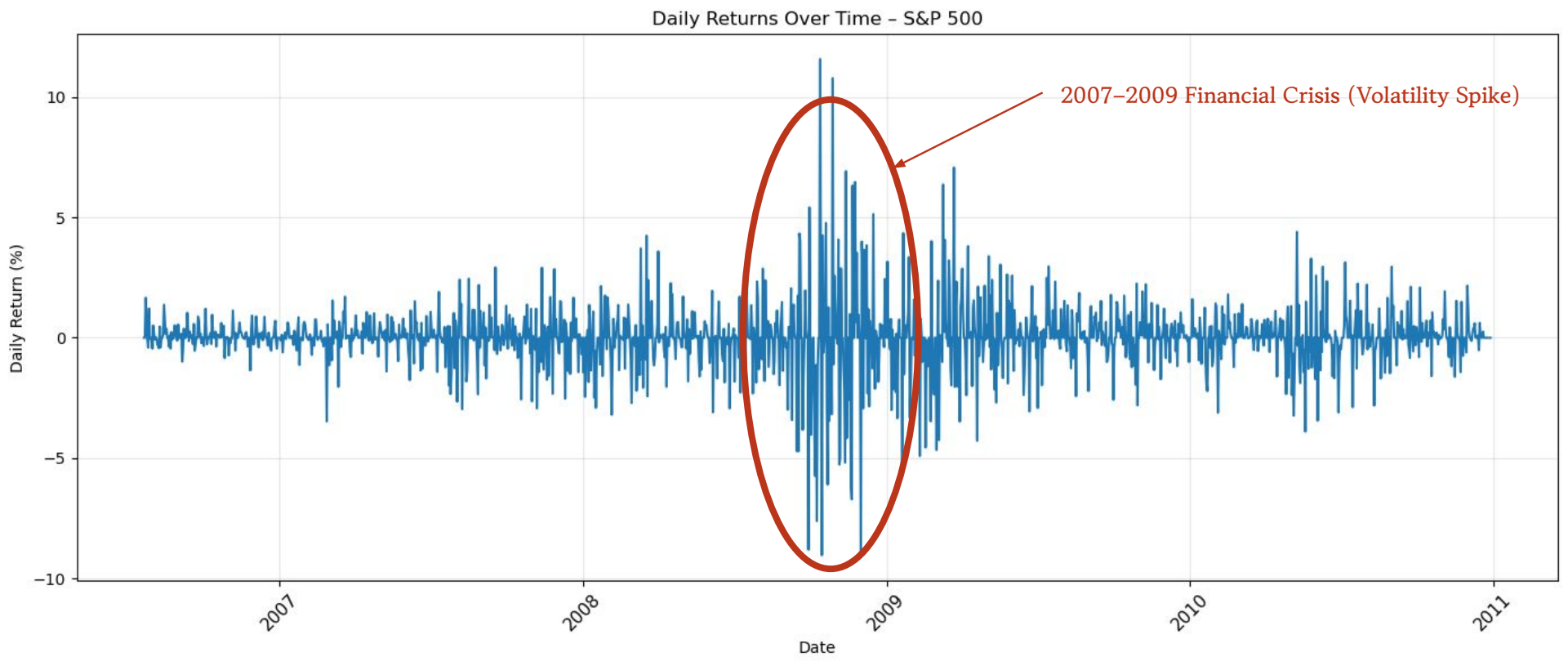
Daily Percent Change Distribution



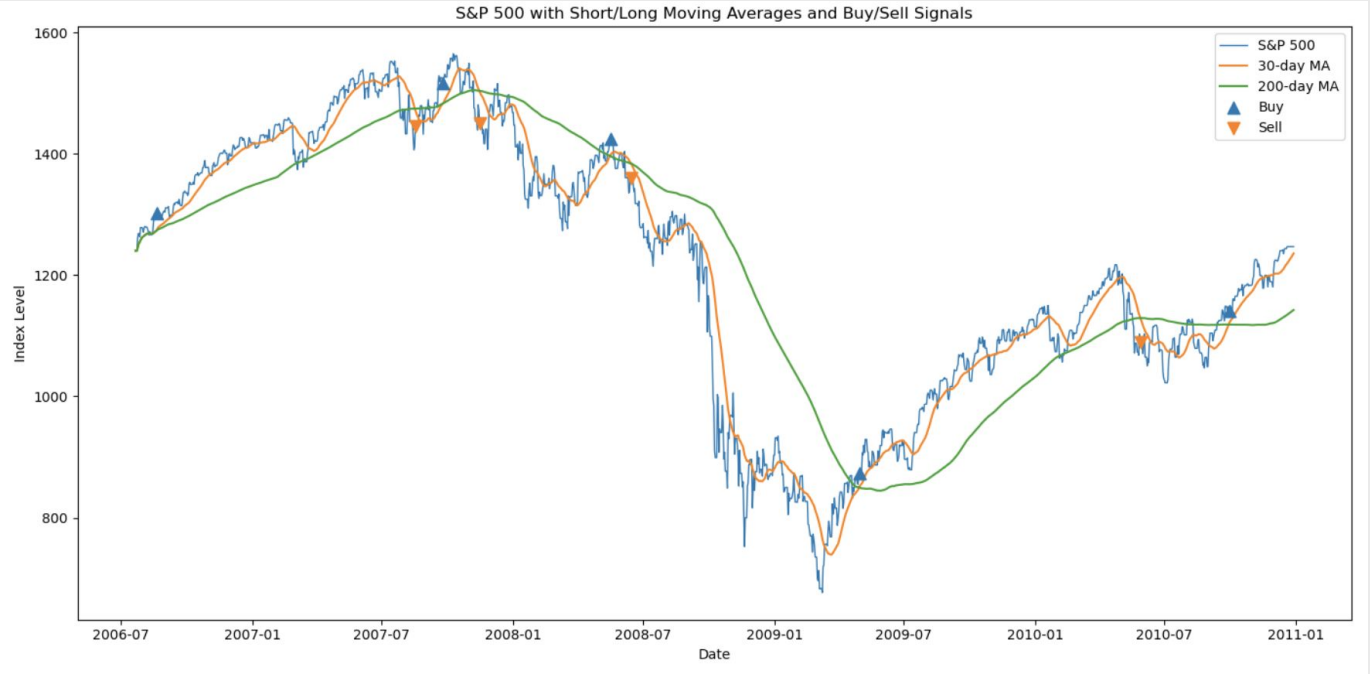


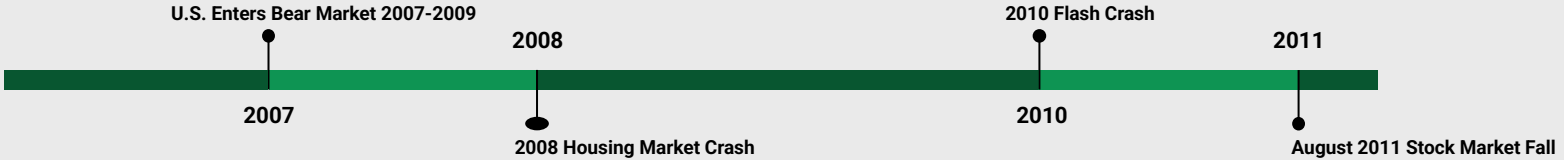
Zoomed-In Distribution

S&P 500 Daily Returns & Volatility (2006–2010)

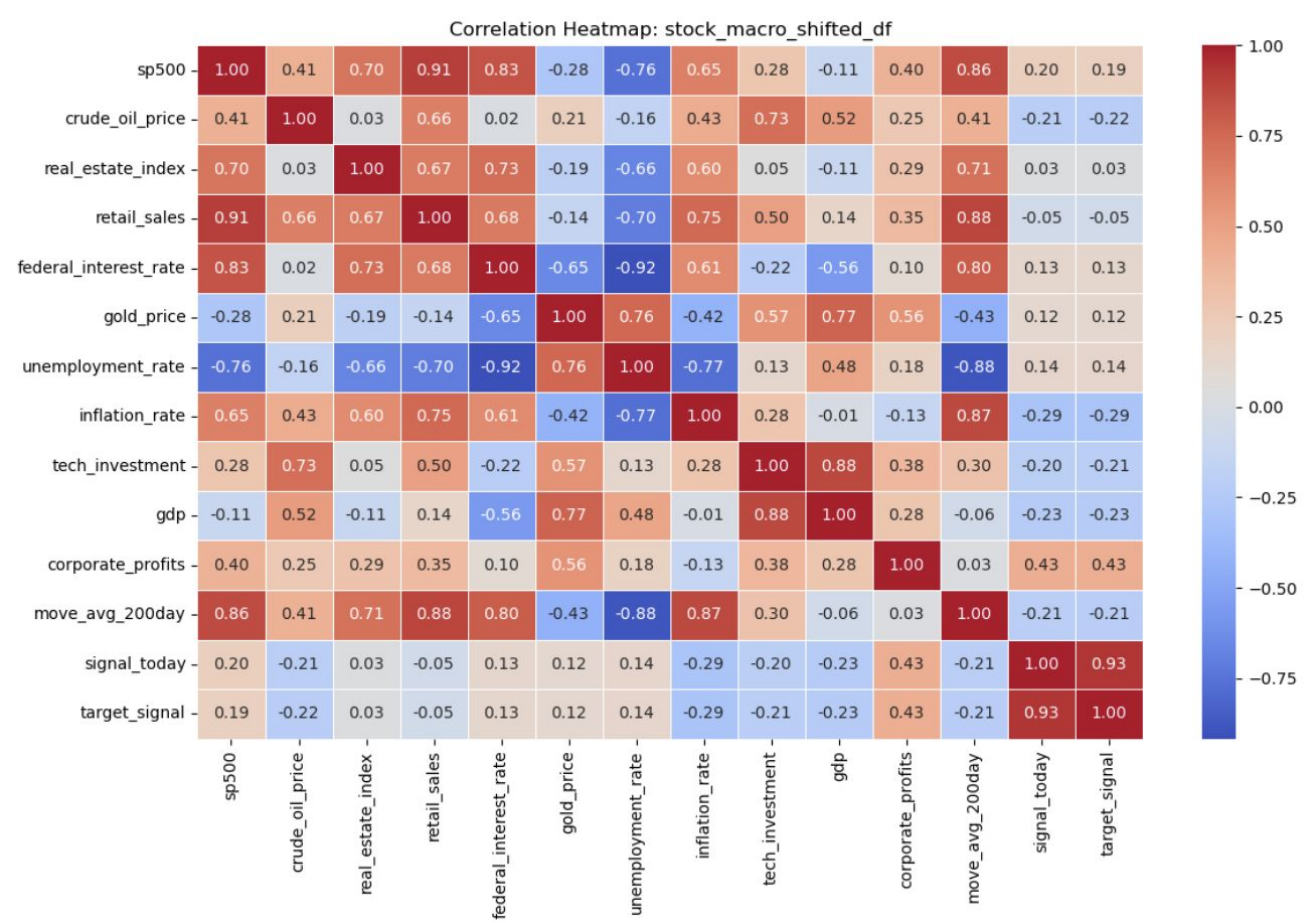


S&P 500 Short/Long Term Moving Averages





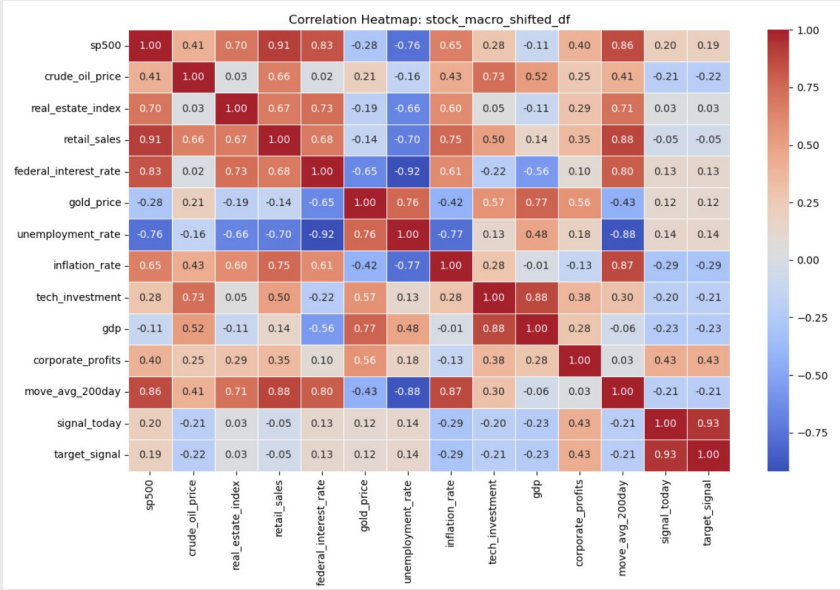
Heat Map


Correlation Heatmap: stock_macro_shifted_df
sp500 1.00 0.41 0.70 0.91 0.83 -0.28 -0.76 0.65 0.28 -0.11 0.40 0.86 0.20 0.19
crude_oil_price 0.41 1.00 0.03 0.66 0.02 0.21 -0.16 0.43 0.73 0.52 0.25 0.41 -0.21 -0.22
real_estate_index 0.70 0.03 1.00 0.67 0.73 -0.19 -0.66 0.60 0.05 -0.11 0.29 0.71 0.03 0.03
retail_sales 0.91 0.66 0.67 1.00 0.68 -0.14 -0.70 0.75 0.50 0.14 0.35 0.88 -0.05 -0.05
federal_interest_rate 0.83 0.02 0.73 0.68 1.00 -0.65 -0.92 0.61 -0.22 -0.56 0.10 0.80 0.13 0.13
gold_price -0.28 0.21 -0.19 -0.14 -0.65 1.00 0.76 -0.42 0.57 0.77 0.56 -0.43 0.12 0.12
unemployment_rate -0.76 -0.16 -0.66 -0.70 -0.92 0.76 1.00 -0.77 0.13 0.48 0.18 -0.88 0.14 0.14
inflation_rate 0.65 0.43 0.60 0.75 0.61 -0.42 -0.77 1.00 0.28 -0.01 -0.13 0.87 -0.29 -0.29
tech_investment 0.28 0.73 0.05 0.50 -0.22 0.57 0.13 0.28 1.00 0.88 0.38 0.30 -0.20 -0.21
gdp -0.11 0.52 -0.11 0.14 -0.56 0.77 0.48 -0.01 0.88 1.00 0.28 -0.06 -0.23 -0.23
corporate_profits 0.40 0.25 0.29 0.35 0.10 0.56 0.18 -0.13 0.38 0.28 1.00 0.03 0.43 0.43
move_avg_200day 0.86 0.41 0.71 0.88 0.80 -0.43 -0.88 0.87 0.30 -0.06 0.03 1.00 -0.21 -0.21
signal_today 0.20 -0.21 0.03 -0.05 0.13 0.12 0.14 -0.29 -0.20 -0.23 0.43 -0.21 1.00 0.93
target_signal 0.19 -0.22 0.03 -0.05 0.13 0.12 0.14 -0.29 -0.21 -0.23 0.43 -0.21 0.93 1.00
sp500
crude_oil_price
real_estate_index
retail_sales
federal_interest_rate
gold_price
unemployment_rate
inflation_rate
tech_investment
gdp
corporate_profits
move_avg_200day
signal_today
target_signal
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75

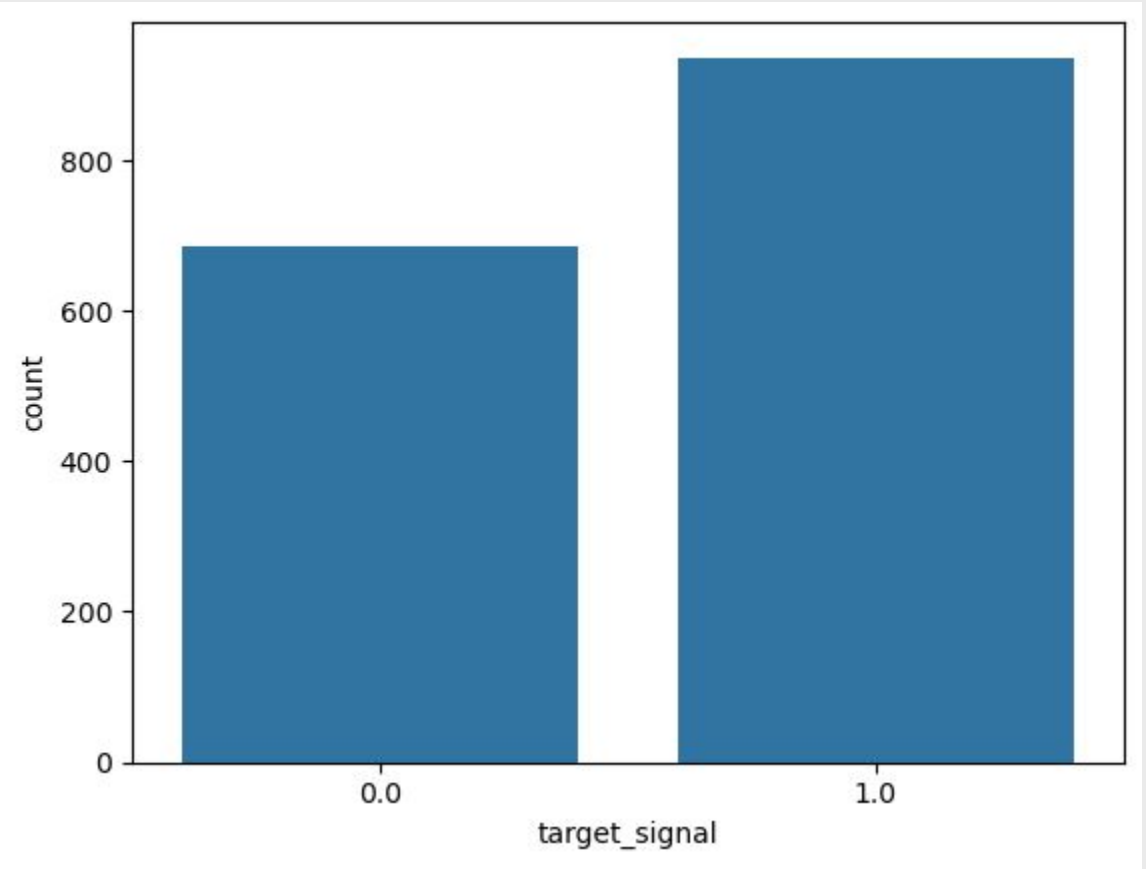
Heat Map Observations

Observations:

- **Strong Positive Correlation :** The S&P 500 has a positive correlation with Retail Sales, Federal Interest Rate, and Tech Investment.
- **Unemployment Counter-Cyclical:** Unemployment Rate is strongly negatively correlated with stocks, spending, real estate, and inflation rate.
- **When to Invest in Gold:** Gold Price is negatively correlated with equities and positively correlated with unemployment, gold tends to rise when markets/economy are weaker.
- **Individual Economic Variable and Signals :** Trading signals have only moderate correlations with any single macro or market variable.



Target Prediction



Trend-Following Strategy:

1 = BUY (Bullish Signal)
Price **above** 200-day moving average -> upward trend

0 = SELL (Bearish Signal)
Price **below** 200-day moving average -> downward trend

200-day MA Signal:

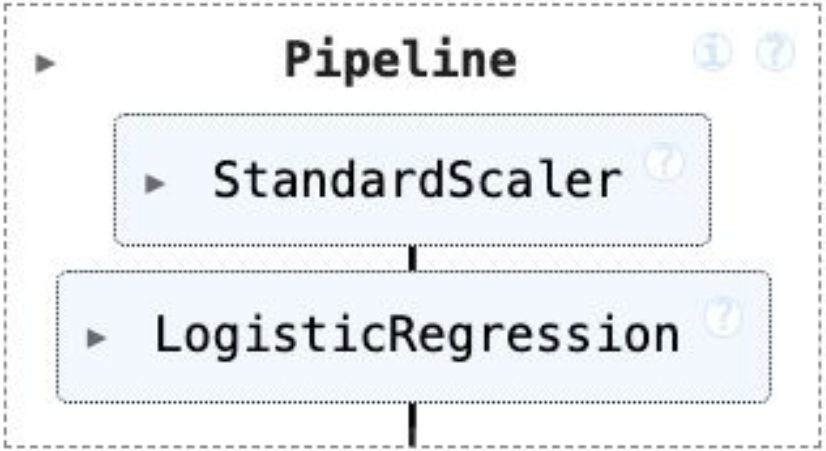
1. Identify long-term trends
2. Reduce risk & avoid major downtrends

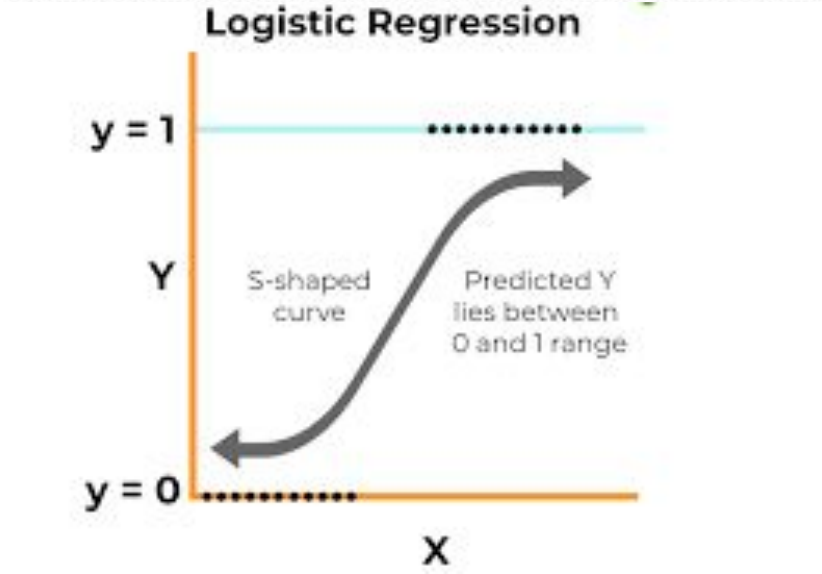
Trend stabilizer, not profit-maximizer

Logistic Regression

Logistic Regression

- **Goal:**
  - Predict when to Buy or Sell
- **Why Logistic?**
  - Simple and Interpretable
  - Built for Binary 0/1 Classification
- **Drawback:**
  - Overfit on high-dimensional dataset
  - Biased on imbalanced data





Classification Report & Confusion Matrix Insights

Overall Accuracy

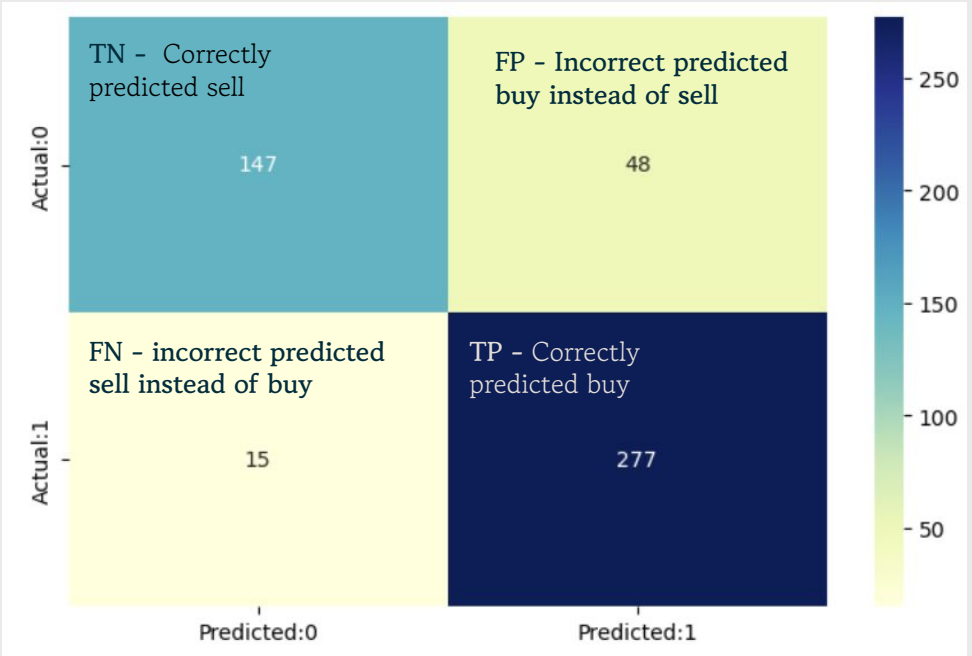
- **87.1% accuracy**
  - *Predicts buy/sell signals correctly 87% of the time.*

Class Performance

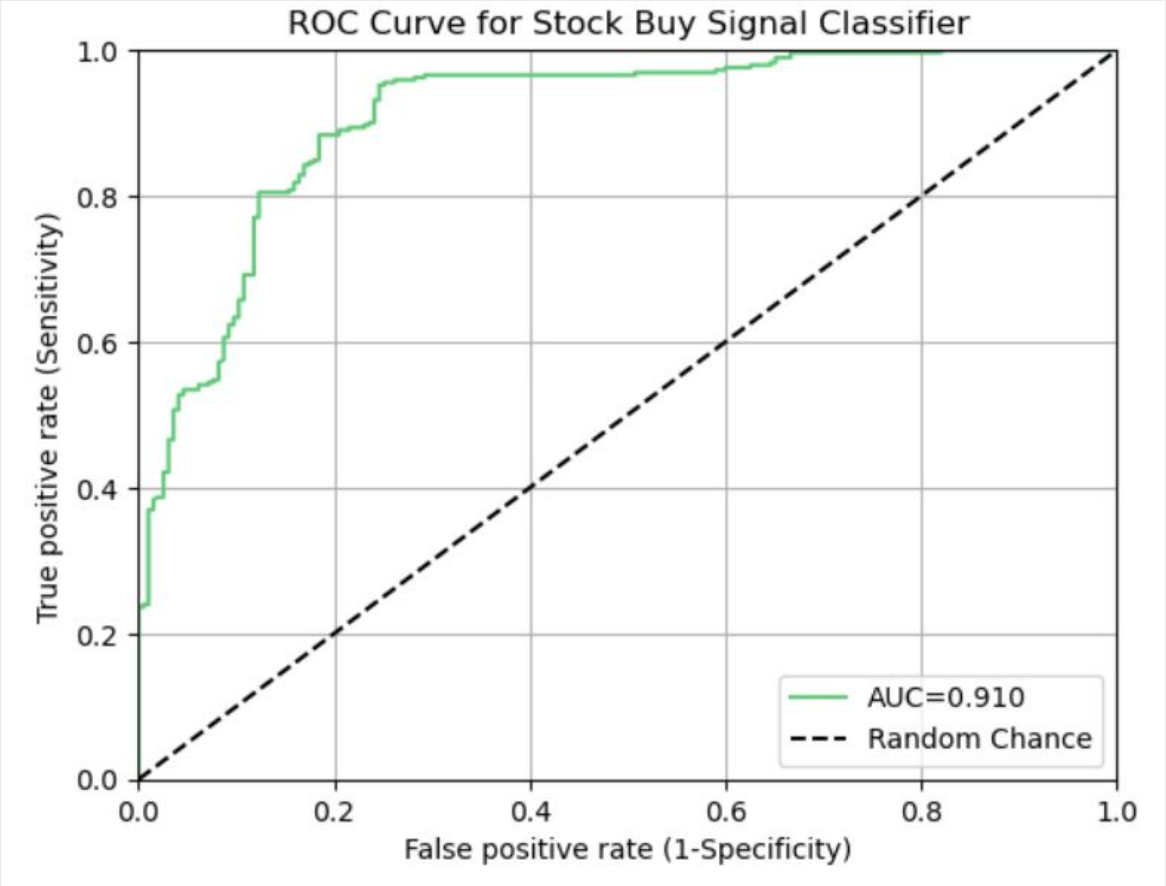
- **Buy (1):** 85% precision, 95% recall
- **Sell (0):** 91% precision, 75% recall
  - *Strong at detecting buy signals, slightly weaker on sell signals.*

Sensitivity & Specificity

- **Sensitivity (TPR): 0.949**
  - *Good at catching "buy" signals*
- **Specificity (TNR): 0.754**
  - *Moderate at identifying "sell" signals*



AUC-ROC Curve



Overall performance of binary classification model

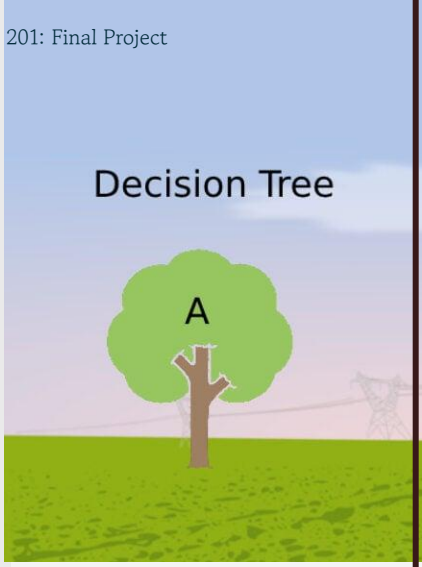
Area Under Curve =
0.910 > 0.5

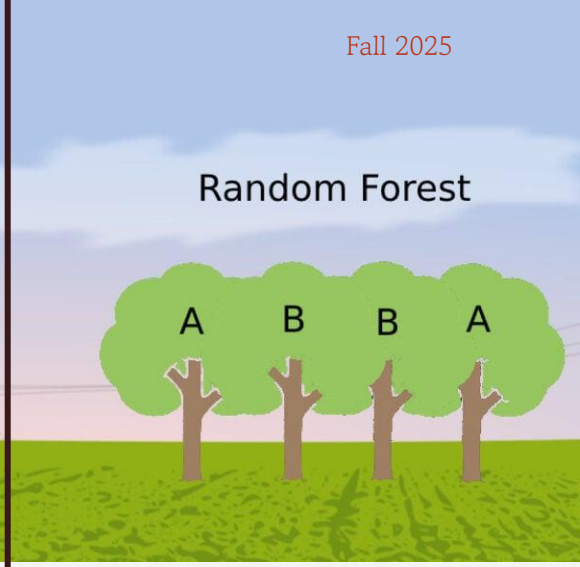
Model performs well & can distinguish accurately between 2 classes

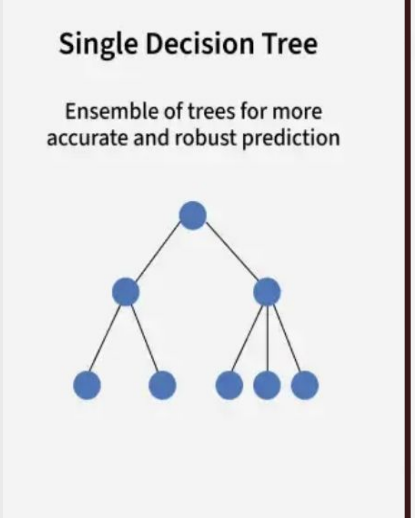
Random Forest

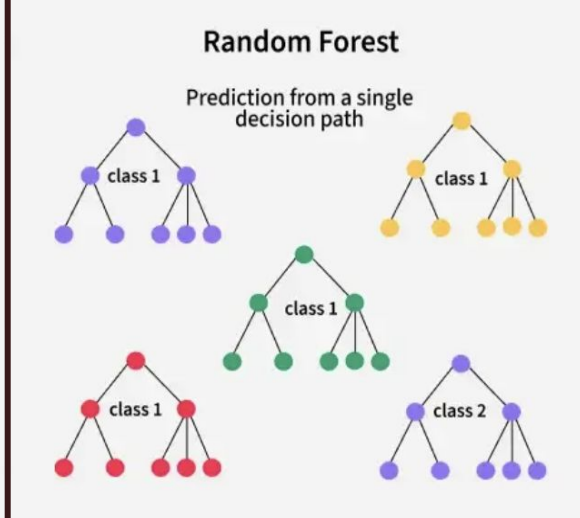
Why Random Forest?

- Goal
  - Predict when to Buy or Sell
- Why Random Forest?
  - Captures non-linear relationships
    - Between macroeconomic indicators & signals
  - Reduce overfitting
    - Averaging many decision trees
  - Provides feature importance
    - Make model more interpretable and insightful
- Drawback:
  - Less interpretable
  - High computational costs

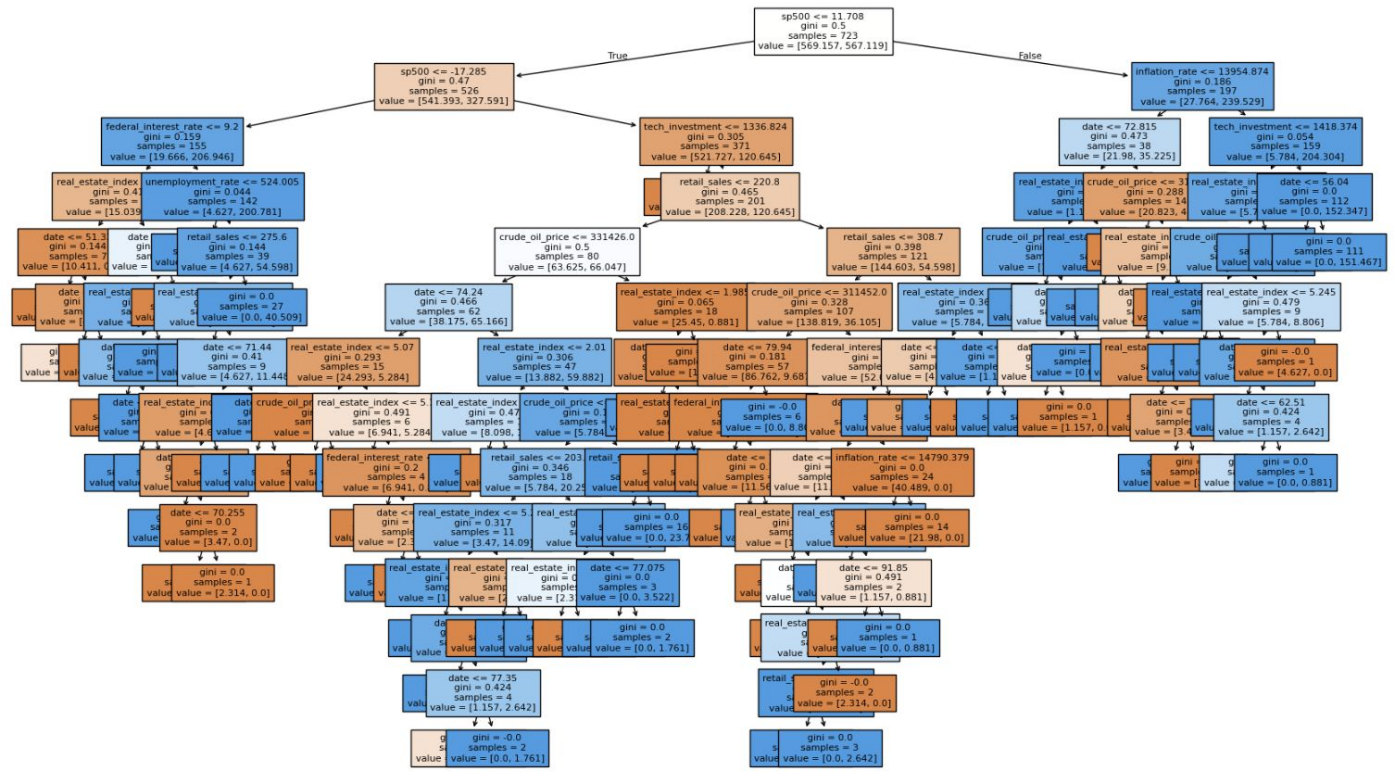








Single Tree in Random Forest



Example decision tree inside the Random Forest

Model uses macroeconomic indicators to classify signals

Forest averages many trees to produce stable, accurate predictions

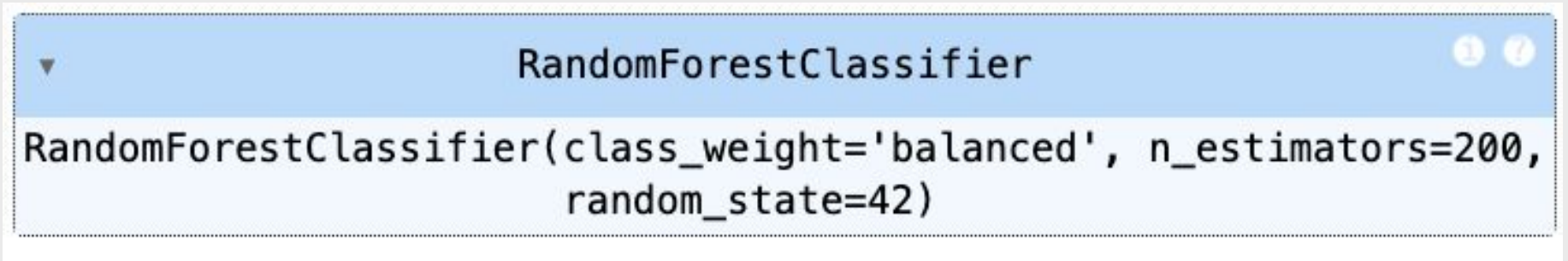
Random Forest Model Performance

Root Impurity

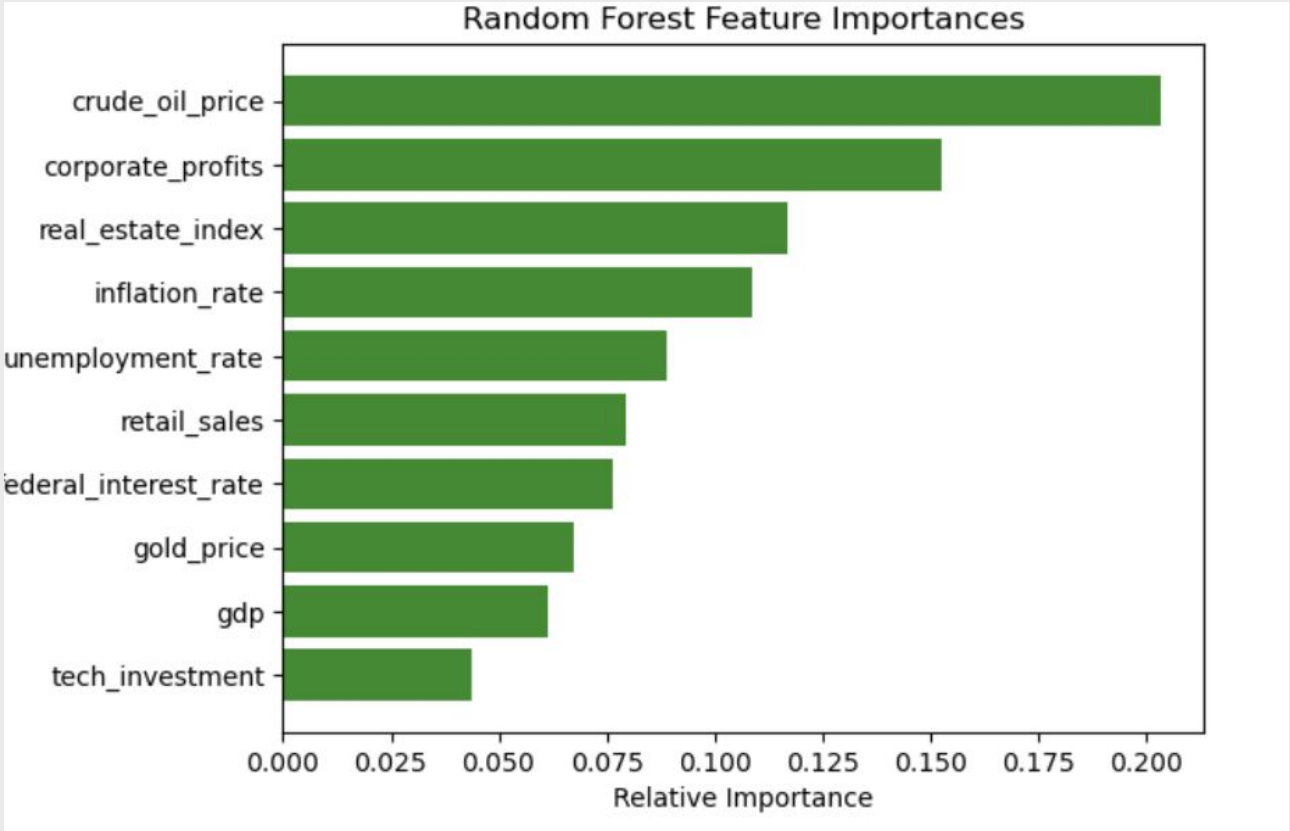
- Average Root Impurity: 0.499
  - Even mix buy/sell distinct classes
  - Model will learn meaningful splits to separate them

Cross-Validation (5-fold)

- Average CV accuracy: 0.751
  - Model predicts buy/signal 75% accuracy on unseen data



Feature Importance

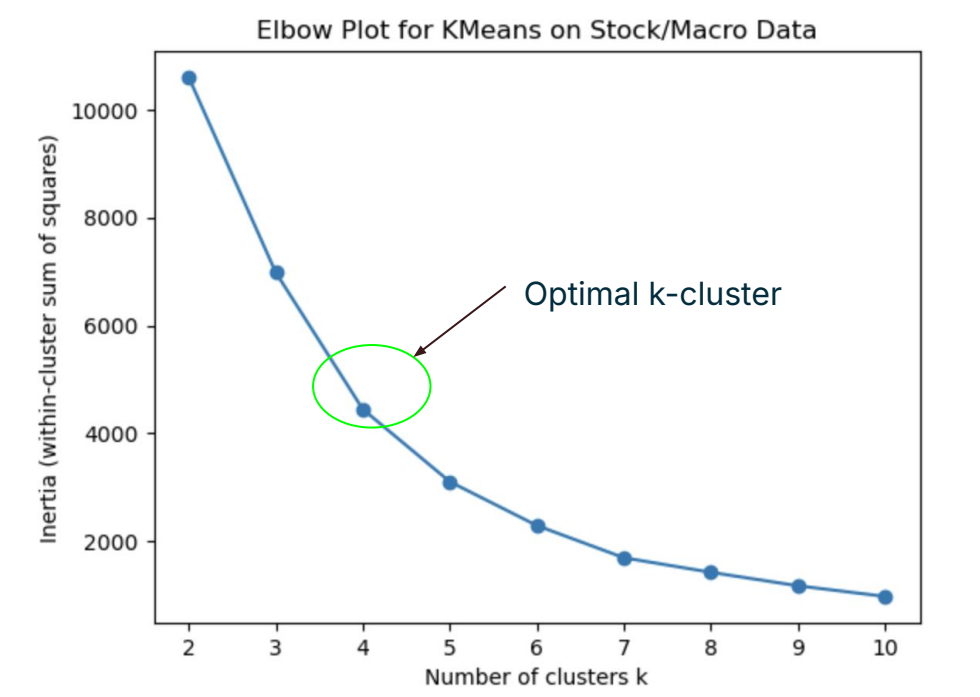


Scores

crude_oil_price	**0.2035**
corporate_profits	**0.1528**
real_estate_index	**0.1171**
inflation_rate	0.109
unemployment_rate	0.0891
retail_sales	0.0795
federal_interest_rate	0.0766
gold_price	0.0674
gdp	0.0612
tech_investment	0.0438

K-Means in PCA Space

Performing K-Means & Picking Optimal k



Within Cluster Sum of Squares:

```
k = 1: WCSS = 19476.000
k = 2: WCSS = 10601.721
k = 3: WCSS = 6983.560
k = 4: WCSS = 4447.645
k = 5: WCSS = 3102.946
k = 6: WCSS = 2290.368
k = 7: WCSS = 1692.535
k = 8: WCSS = 1422.432
k = 9: WCSS = 1172.657
k = 10: WCSS = 975.343
```

Cluster Sizes at k=4:

```
Cluster sizes:
cluster_kmeans
1.0    455
0.0    437
2.0    367
3.0    364
```

Transforming K-Means Data into 2D PCA Space

Principal Component 1 Top Features

Top features for PC1:

	PC1
move_avg_200day	0.392238
unemployment_rate	−0.379046
sp500	0.372836
federal_interest_rate	0.366489
retail_sales	0.364851
inflation_rate	0.344086
real_estate_index	0.314696
gold_price	−0.205934
crude_oil_price	0.148881
gdp	−0.090078
tech_investment	0.062971
corporate_profits	0.044207

Principal Component 2 Top Features

Top features for PC2:

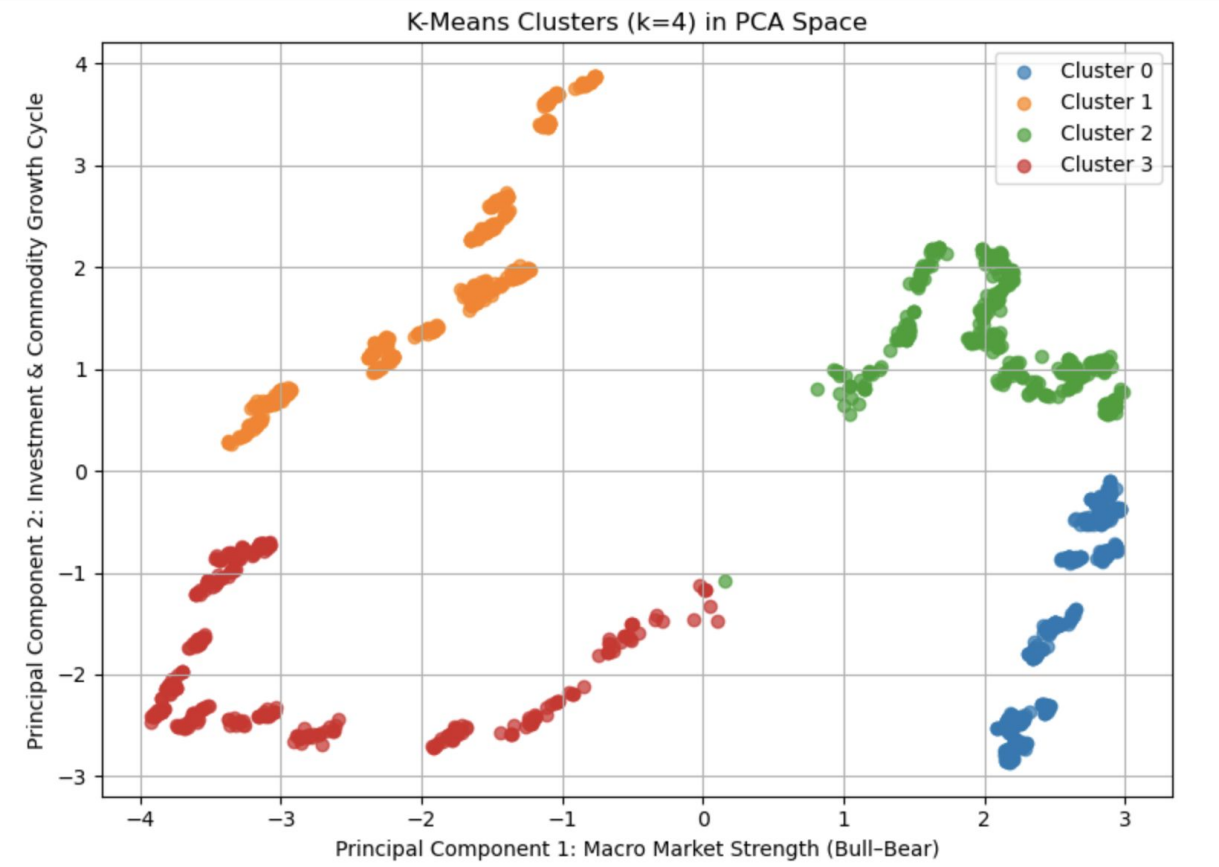
	PC2
tech_investment	0.496256
gdp	0.480196
gold_price	0.407907
crude_oil_price	0.378043
corporate_profits	0.300889
retail_sales	0.216297
federal_interest_rate	−0.177545
unemployment_rate	0.172856
sp500	0.098597
move_avg_200day	0.054121
inflation_rate	0.048467
real_estate_index	0.010896

Explained Variance

PC1: 0.501
PC2: 0.297
PC3: 0.110
PC4: 0.050

PC1 and PC2 capture about 80% of the variation for the 12 macroeconomic/market variables.

K-Means Clusters in PCA Space



Target Signal Distribution by Cluster

Cluster 0: Target Signal usually says 1 in this healthy-bull Market

Cluster 1:Still mostly **1**, but less extreme than cluster 0.

Cluster 2: In the high-inflation market, the signal is **usually 0**.

Cluster 3:In the stressed/bear market, the signal is mixed, slightly more 0 than 1.

Target signal distribution by cluster:

cluster_kmeans	target_signal	
0.0	1.0	0.846682
	0.0	0.153318
1.0	1.0	0.740659
	0.0	0.259341
2.0	0.0	0.822888
	1.0	0.177112
3.0	0.0	0.546703
	1.0	0.453297

Cluster Features and Means

Cluster 0: A **healthy bullish regime**: strong market, low unemployment, normal-ish rates & inflation. • **High S&P 500** (~1425) • **Moderate interest rate** (~5.2) • **Low unemployment** (~4.5) • **Moderate inflation** (~3.0)	Cluster 1: A **post-crisis environment**: weak labor market but super-low inflation rates and supportive policy. • **S&P a bit lower price** (~1128) • **High unemployment** (~9.7) • **Low inflation** (~1.2) • **Highest GDP** & corporate profits, high gold price
Cluster 2: A **high-inflation & high-oil boom regime**: there growth but inflationary. • **High S&P** (~1367) • **Very high oil price** (~105) • **Highest retail sales & tech investment** • **Highest inflation** (~3.6), Mid-level unemployment	Cluster 3: A **stressed / bear regime**: weak markets, weak profits, high joblessness. • **Lowest S&P** (~902) • **Very weak real estate** (most negative real_estate_index) • **High unemployment** (~8.5), **Very low inflation** (~0.7) • **Lowest corporate profits, Lowest 200-day moving average**

Model Comparison

Model Accuracy Comparison

Logistic Regression	Random Forest (Avg 5-Fold Cross Validation)	K-Means in PCA Space
0.871	0.7511	76.8% Explained Variance

- **Logistic Regression**
  - Strong, reliable baseline model for classification tasks
  - Simple, fast, interpretable
  - Outperformed complex model
- **Random Forest**
  - More complex and powerful
  - Lower accuracy, may overfit
- **K-means in PCA Space**
  - No accuracy, deterministic
  - Model captures about 80% of variance in data

Conclusions, Limitations, & Future Improvements

Conclusions

Exploratory Analysis

- Identified economic/market growth and crashes reflected in the S&P 500
- S&P 500 Correlation:
  - **Positively** related to retail sales, federal interest rate, and tech investment.
  - **Negatively** related to unemployment rate, and (typically) gold price
    - Gold price tends to rise when market and economy are weak

Statistical Modeling

- **Logistic Regression:** Achieved the **highest accuracy at about 87%** and **strong AUC (0.91)**.
  - This gives reliable baseline for predicting Buy (1) vs Sell (0) signals.
- **Random Forest:** Captured **non-linear relationships** but **averaged at about 75% CV accuracy**, lower than logistic.
  - Indicating the buy/sell boundary is **generally smooth and linear**, so not highly complex.
- **K-Means in 2D PCA Space (k=4):** Found four meaningful market regimes (0. Healthy/Bull Market, 1. Post-Crisis/Low Interest Rates, 2. High-Inflation/Oil Boom, and 3. Stressed/Bear Market) with different buy/sell signal frequencies.

Importance of Economic/Market Analysis

- The relationship between economic factors and market trends can inform prevention tactics to prevent future market crashes.

Limitations & Future Improvements

Limitations:

- **Limited Timeframe:** Focuses on years 2006-2010 in the United States, and heavily influenced by 2008 market crash.
  - Models may *not generalize* to other periods of the S&P 500, or to other countries/regions economy.
- **Macroeconomic Data is Forward-Filled:** Some economic factors are reported monthly (e.g. GDP, unemployment rates, etc.). Due to the low-frequency of particular variables must forward fill data to match data points
  - May result in mismatched timing and extra noise
- **Simplified Trading (Target) Signal:** Does not consider transaction costs, position sizing, and actual return amount.

Future Improvements:

- **Broaden Dataset:** add more years, compare different stocks indices, investigate other economic indicators
- **Expanded Target Signals and Performance:** Create multi-class signals (e.g. strong buy, call, put, hold etc.)
- **Random Forest Pruning Hyperparameter Tuning:** maximum features, maximum depth, minimum samples lead, etc.

Thank You!

Any Comments or Questions?