Predicting Economic Signals: Modeling Core PCE Inflation and the S&P 500 Index

Introduction

Understanding and forecasting key economic indicators is central to both macroeconomic policymaking and financial decision-making. This project focuses on building predictive models for two crucial targets in the U.S. economy: Core Personal Consumption Expenditures (Core PCE) inflation and the S&P 500 index. The Core PCE serves as the Federal Reserve's preferred inflation gauge, directly influencing monetary policy decisions such as interest rate adjustments. Meanwhile, the S&P 500 index reflects market sentiment and expectations regarding the broader economy, and is a key benchmark for equity market performance.

Leveraging a comprehensive dataset combining macroeconomic indicators, labor market data, interest rates, commodity prices, and policy variables (such as average tariff rates), we develop machine learning models to understand and predict these targets. In the first part of the project, we model Core PCE inflation using a wide array of real and nominal variables to examine the structural drivers of inflationary pressure. In the second part, we shift focus to the S&P 500, constructing a linear regression model that isolates the effect of macroeconomic fundamentals—excluding market-specific variables—to evaluate how economic conditions relate to equity market movements.

Our models employ scikit-learn pipelines with preprocessing steps such as standardization and one-hot encoding, allowing for efficient feature engineering and robust performance evaluation. The results highlight the most influential variables for each predictive task and provide interpretability through coefficient analysis in the linear model.

This dual-focus approach provides not only forecasting utility but also economic insight, bridging macro-level fundamentals with real-time policy and market signals.

The dataset includes a comprehensive array of macroeconomic, financial, and policy-related variables that are critical to understanding both inflation dynamics and equity market movements. The (unemployment_rate) reflects the proportion of the labor force that is unemployed and actively seeking work, serving as a key indicator of labor market conditions. The (cpi_all_items) captures overall inflation including food and energy, while the (core_pce) excludes these volatile components and serves as the Federal Reserve’s primary inflation gauge. Measures of output and income include (real_gdp) and (nominal_gdp), which track inflation-adjusted and current-dollar economic production respectively. The (federal_funds_rate) denotes the interest rate at which depository institutions lend to each other overnight and is a primary tool

of U.S. monetary policy. Long- and short-term interest rates are represented by the (10y_treasury_yield) and (3m_treasury_yield), capturing yield curve dynamics. The (m2_money_stock) represents the broad money supply, while labor market activity is further detailed by (initial_jobless_claims) and (continuing_claims), which track unemployment insurance filings.

Wage pressures and labor income trends are captured through (average_hourly_earnings) and (employment_total_nonfarm), indicating employment levels across sectors. Housing market conditions are reflected in (case_shiller_us) and (house_price_index), as well as (housing_starts), which measure new residential construction. Consumption and investment in durable goods are proxied by (total_vehicle_sales). Commodity price movements are represented by (wti_crude_oil) and (brent_crude_oil), both of which impact inflation expectations and producer costs. The (us_dollar_index) measures the value of the U.S. dollar against major foreign currencies, while exchange rate variables such as (usd_jpy), (usd_eur), and (usd_cny) provide insight into international purchasing power and trade dynamics.

Market risk sentiment is proxied by the (vix_index), a volatility index, while (pce_inflation) represents overall personal consumption expenditure price trends. Measures of consumer and business activity include (personal_income), (retail_sales), (industrial_production), and (manufacturing_ip). Capacity constraints are captured through (capacity_utilization), and credit market behavior is reflected in (business_loans) and (credit_card_loans). Consumer expectations are addressed via (consumer_sentiment). The transportation sector is detailed through (truck_employment), (truck_tonnage_index), (heavy_truck_sales), and (truck_production_ppi), which indicate logistics intensity and industrial demand. Additional inflation-related indicators include (used_cars_cpi) and (corporate_bond_yield_baa), the latter reflecting credit risk premiums. Measures of real income and consumption such as (real_disposable_income) and (real_personal_consumption) further contextualize household behavior. The (personal_savings_rate) indicates consumer caution or confidence, while (st_louis_fin_stress) tracks financial system stress. Finally, (core_inflation_services) highlights service-sector price movements, and monetary conditions are additionally captured by variables such as (prime_rate), (10y_minus_2y), (10y_minus_3m), (federal_funds_effective), (overnight_rp), and (10y_nominal), all of which reflect the prevailing interest rate environment and expectations about future economic conditions.

In addition to macroeconomic and monetary indicators, the dataset includes variables capturing equity markets, commodity markets, digital assets, and policy instruments. The (sp500) represents the S&P 500 index, a benchmark for the performance of large-cap U.S. equities, while (nasdaq) and (dow_jones) offer complementary views of technology-heavy and industrial stock performance, respectively. The (recession_flag) is a binary indicator that marks periods of official U.S. economic recessions, serving as a regime variable in temporal modeling. A set of individual stock prices—such as (AAPL), (MSFT), (GOOGL), (META), (AMZN), (NVDA),

(TSLA), and others—represents firm-level equity trends and is primarily used in exploratory analysis or excluded from fundamental-only models.

In the realm of commodities, the dataset includes (GC=F) for gold prices, (CL=F) for crude oil, (NG=F) for natural gas, and several agricultural futures: (ZW=F) for wheat, (ZC=F) for corn, and (ZS=F) for soybeans. These prices are relevant for analyzing supply-side inflation shocks and cost-push pressures. Digital asset indicators such as (BTC-USD) for Bitcoin, (ETH-USD) for Ethereum, and (SOL-USD), (BNB-USD), (DOGE-USD) represent alternative investment activity and investor sentiment, although their volatility may limit predictive value in macro models. Stock prices of major corporations including (XOM), (GE), (IBM), (JNJ), (PG), (WMT), (KO), (INTC), (ORCL), and (CVX) provide additional microeconomic indicators that may reflect sector-specific shocks or broader market trends.

Lastly, the (avg_tariff_rate) captures the average effective tariff level imposed by the U.S., serving as a proxy for trade policy stance. Rising tariffs may indicate protectionist shifts that affect both consumer prices and global supply chains, and thus this variable is especially important in modeling cost-driven inflation under scenarios involving trade shocks.

Models and Methods

This project involves two major predictive modeling tasks: forecasting the S&P 500 index and forecasting Core PCE inflation, both using macroeconomic, monetary, and financial variables. The methodology applies a combination of linear and non-linear machine learning models, as well as multiple experimental design schemes to compare economic hypothesis structures.

1. Data Preparation and Feature Engineering

All models were trained on a cleaned and merged macro-financial dataset containing 97 features from 1980 to the present. For both tasks, missing values were removed, and temporal structure was preserved (i.e., no shuffling of rows). A quarter categorical variable was engineered from the timestamp and one-hot encoded to account for seasonal effects. Features for the S&P 500 model excluded contemporaneous market indicators to avoid look-ahead bias, while Core PCE prediction used multiple thematic subsets of variables.

2. Linear Regression Pipeline

As a baseline model, Linear Regression was implemented using a scikit-learn pipeline. Numerical variables were standardized, and the quarter variable was one-hot encoded. This model was used in both tasks to establish an interpretable benchmark, where coefficients could be directly analyzed to infer the direction and magnitude of each macroeconomic driver.

3. Random Forest Regressor

A RandomForestRegressor was used extensively to capture non-linear relationships and interaction effects. It was applied to both:

- S&P 500 prediction, as part of a model comparison alongside other regressors.
- Core PCE prediction, across four economic scheme-based feature sets:
    - Scheme 1 (Monetary variables): interest rates, M2, central bank policy indicators.
    - Scheme 2 (Demand-side variables): GDP, income, consumption, employment.
    - Scheme 3 (Supply-side variables): production, input costs, industrial capacity.
    - Scheme 4 (Financial/asset variables): equity indices, commodity prices, credit risk.

This allowed for direct testing of competing macroeconomic theories using a consistent modeling framework.

4. Decision Tree Regressor

The DecisionTreeRegressor was employed to explore model behavior in a highly interpretable yet flexible form. While its low bias makes it useful for identifying structure in the data, its high variance resulted in significant overfitting for small training sizes, as shown in the learning curves.

5. K-Nearest Neighbors (KNN)

The KNeighborsRegressor served as a non-parametric benchmark. It predicts based on similarity to historical macroeconomic conditions. Although its performance improves with more data, its sensitivity to local noise limits its utility in high-dimensional macroeconomic forecasting.

6. Baseline Regressor

A simple baseline model was included that predicts the mean of the target variable. This naive predictor provided a reference point to evaluate the added value of machine learning models.

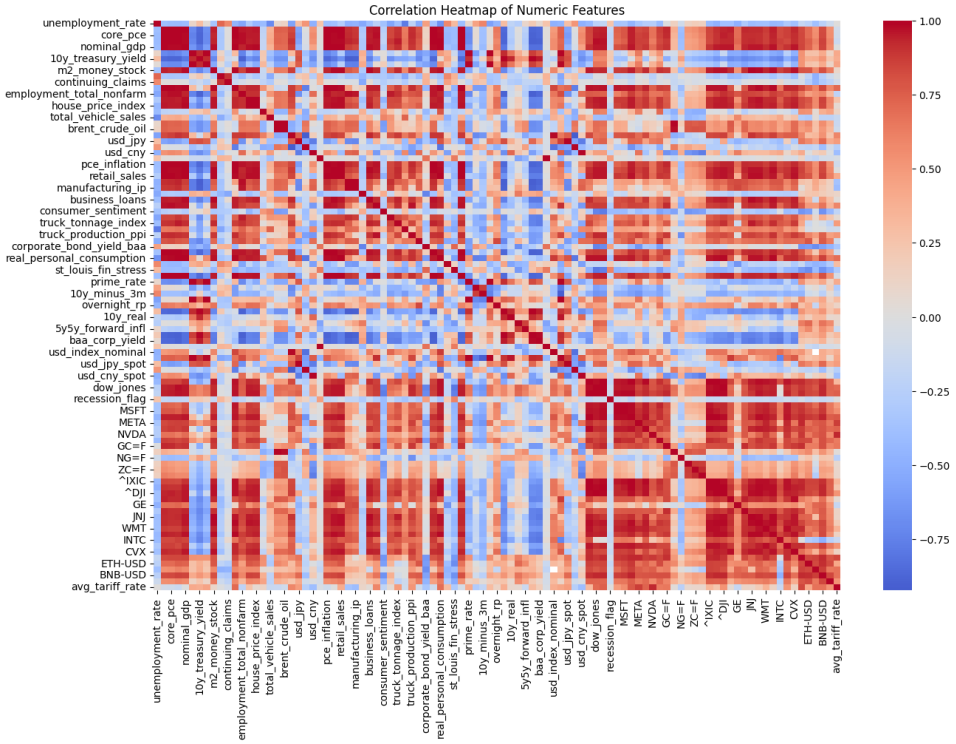
7. Model Evaluation

Each model was evaluated using:

- Mean Squared Error (MSE) on both training and validation sets
- R-squared (R^2) for explanatory power
- Learning Curves, which visualized model performance across increasing training set sizes and helped assess overfitting versus underfitting tendencies.

For S&P 500, Random Forest achieved the best validation performance, closely followed by Linear Regression. For Core PCE, model performance varied across schemes, with the Monetary and Financial schemes showing relatively better alignment with actual inflation trends.

Results and Interpretation



Correlation Analysis of Macro-Financial Variables

To better understand the linear relationships across the dataset, a Pearson correlation heatmap was generated for all numerical features. The heatmap (see Figure X) reveals several intuitive and notable patterns:

- The (core_pce) is positively correlated with (nominal_gdp), (pce_inflation), and (retail_sales), confirming the strong link between inflation and nominal demand indicators.
- (unemployment_rate) is negatively correlated with (core_pce) and (sp500), reflecting classic Phillips Curve and business cycle dynamics.
- Monetary indicators such as (federal_funds_rate), (10y_treasury_yield), and (m2_money_stock) exhibit moderate correlations with inflation and asset price variables, supporting their inclusion in the monetary scheme of Core PCE modeling.

- Notably, cryptocurrency prices like (BTC-USD), (ETH-USD) and equity indices such as (NASDAQ) and (S&P 500) are highly correlated among themselves, but much less so with fundamental macro variables, which aligns with their exclusion from macro-only prediction models.

This correlation matrix helped inform feature grouping for scheme-based model comparisons, and validated economic expectations about the interdependence between inflation, interest rates, output, and asset prices.

Interpretation of Linear Regression Coefficients for S&P 500

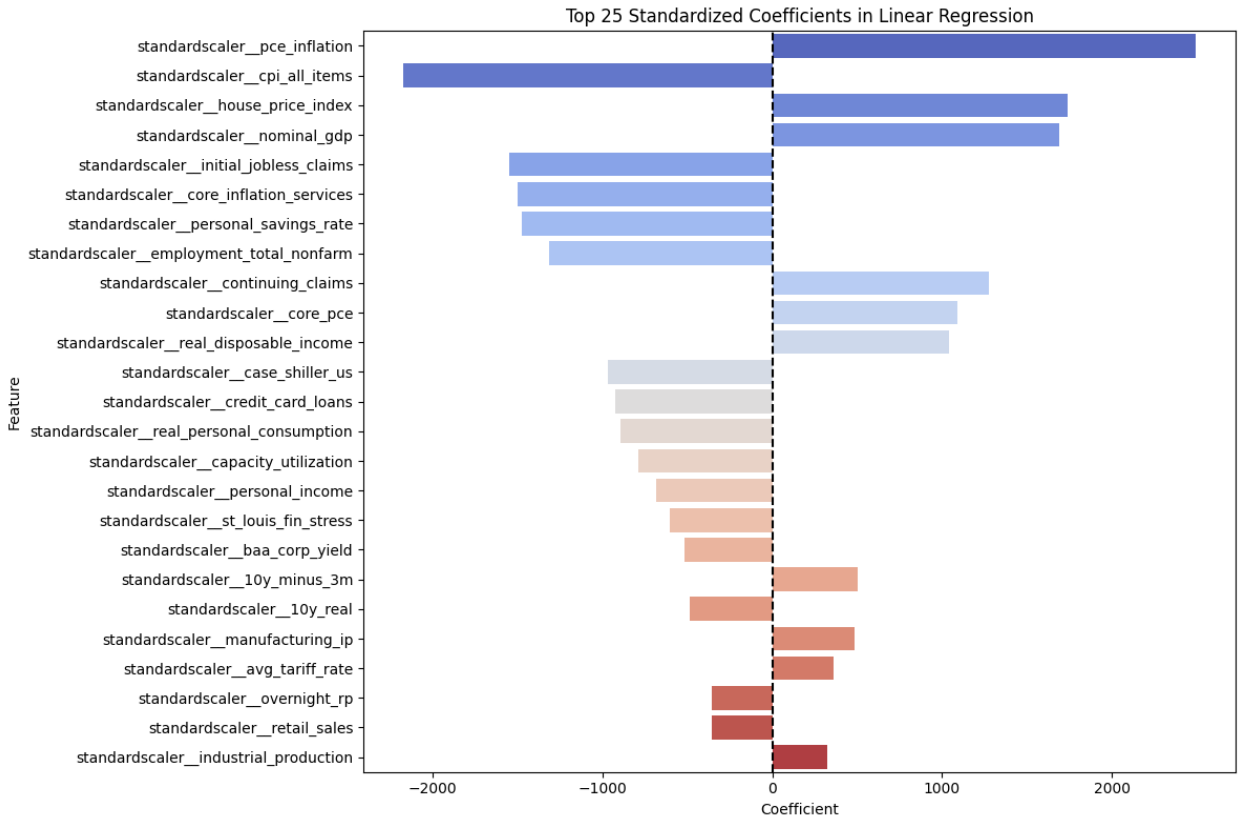
To further interpret the economic drivers of the S&P 500 index, a linear regression model was trained using standardized macroeconomic variables. The model achieved an excellent fit with an R^2 of 0.9998 on the test set and very low MSE (94.2), suggesting a strong linear relationship between the selected economic fundamentals and stock market performance.

As shown in the standardized coefficient plot (Figure X), the most influential positive predictors of the S&P 500 index include (pce_inflation) and (house_price_index), indicating that rising inflation expectations and real asset prices are associated with equity growth. (nominal_gdp) and (continuing_claims) also show strong positive contributions, reflecting the pro-cyclicality of market performance with economic expansion and labor market resilience.

Conversely, variables like (cpi_all_items), (initial_jobless_claims), and (core_inflation_services) hold large negative coefficients, suggesting that certain inflationary pressures and labor market slack may be negatively perceived by equity markets. Notably, (personal_savings_rate) appears as a significant negative predictor, possibly reflecting reduced consumer spending and investment sentiment.

The inclusion of the (avg_tariff_rate) variable yielded a modest negative coefficient, aligning with the hypothesis that higher trade barriers may dampen market valuations due to increased input costs and supply chain frictions.

Overall, this model not only provided high predictive accuracy but also generated economically interpretable insights consistent with macro-financial theory.



Interpreting Feature Importance in the K-Nearest Neighbors (KNN) Model

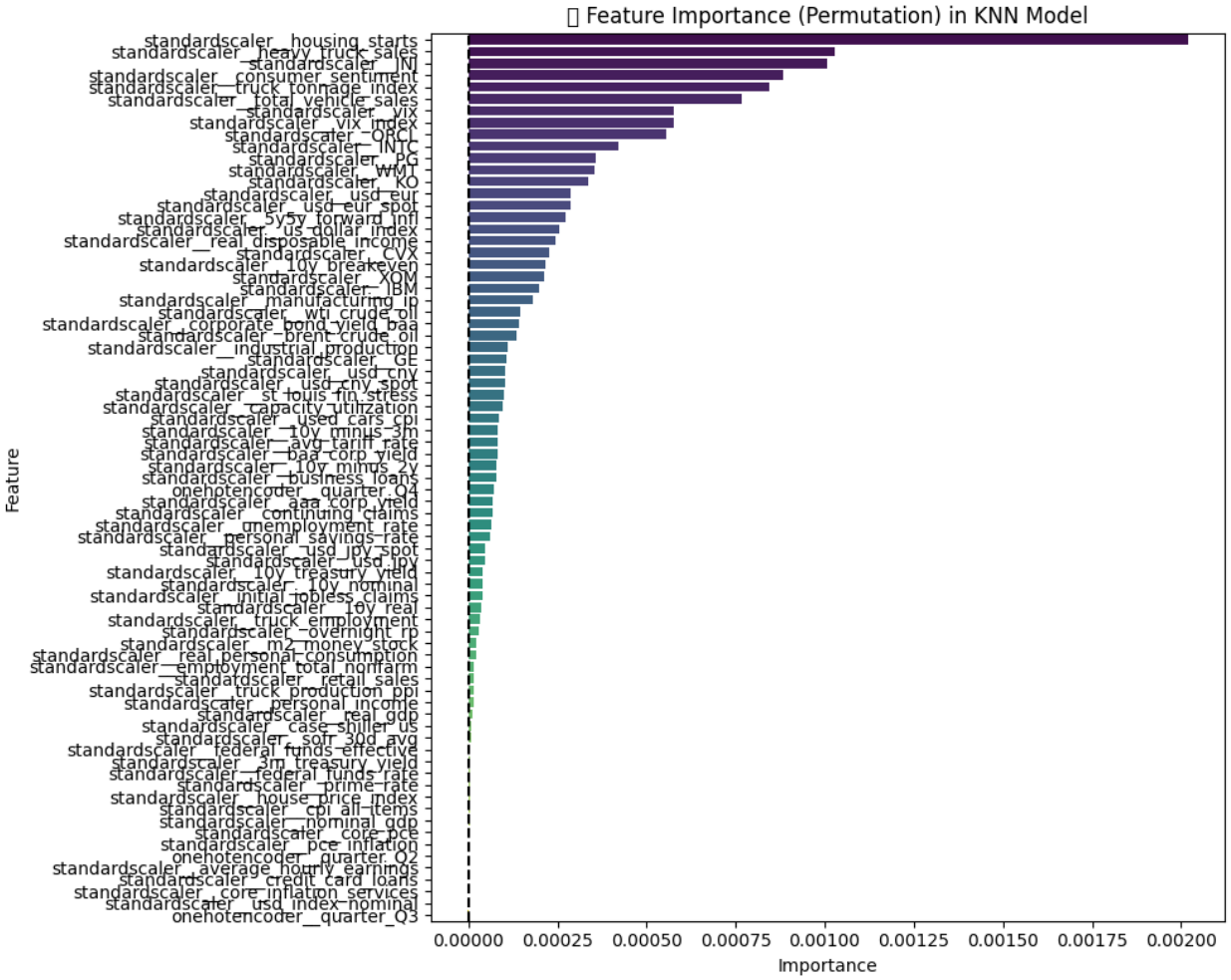
Given that K-Nearest Neighbors (KNN) is a non-parametric model, it does not naturally yield interpretable coefficients. To overcome this limitation, Permutation Importance was employed to quantify the impact of each standardized feature on the model's prediction accuracy.

The resulting importance plot (Figure X) reveals that variables related to construction activity and transportation logistics were among the most influential. Specifically, (housing_starts), (heavy_truck_sales), and (truck_tonnage_index) ranked highest, suggesting that real activity in goods production and infrastructure strongly correlates with market behavior under local macroeconomic conditions. Additionally, (consumer_sentiment) and select corporate indicators such as (JNJ) and (IBM) also showed high predictive contribution in the KNN framework.

Conversely, some traditionally important macroeconomic variables—such as (core_inflation_services), (average_hourly_earnings), and (credit_card_loans)—registered low or even negligible permutation scores in this context. This suggests that while these variables may be valuable in parametric models, they contribute less to distance-based local pattern matching as used in KNN.

Interestingly, the (quarter) dummy variable had near-zero or even slightly negative importance, indicating that seasonality offered limited improvement to predictive performance in a non-parametric setting.

These findings demonstrate that even black-box-like models can be unpacked with appropriate tools, offering nuanced insight into which features most influence predictions through local similarity rather than global correlation or linear influence.



Random Forest Regression: Feature Importance and Performance

To complement the linear and KNN models, a Random Forest Regressor was trained on macroeconomic variables to predict the S&P 500 index. The pipeline included one-hot encoding of quarterly seasonality and standardization of all numerical features. A grid search over tree depth hyperparameters was conducted, yielding an optimal configuration with 100 estimators and a max depth of 10.

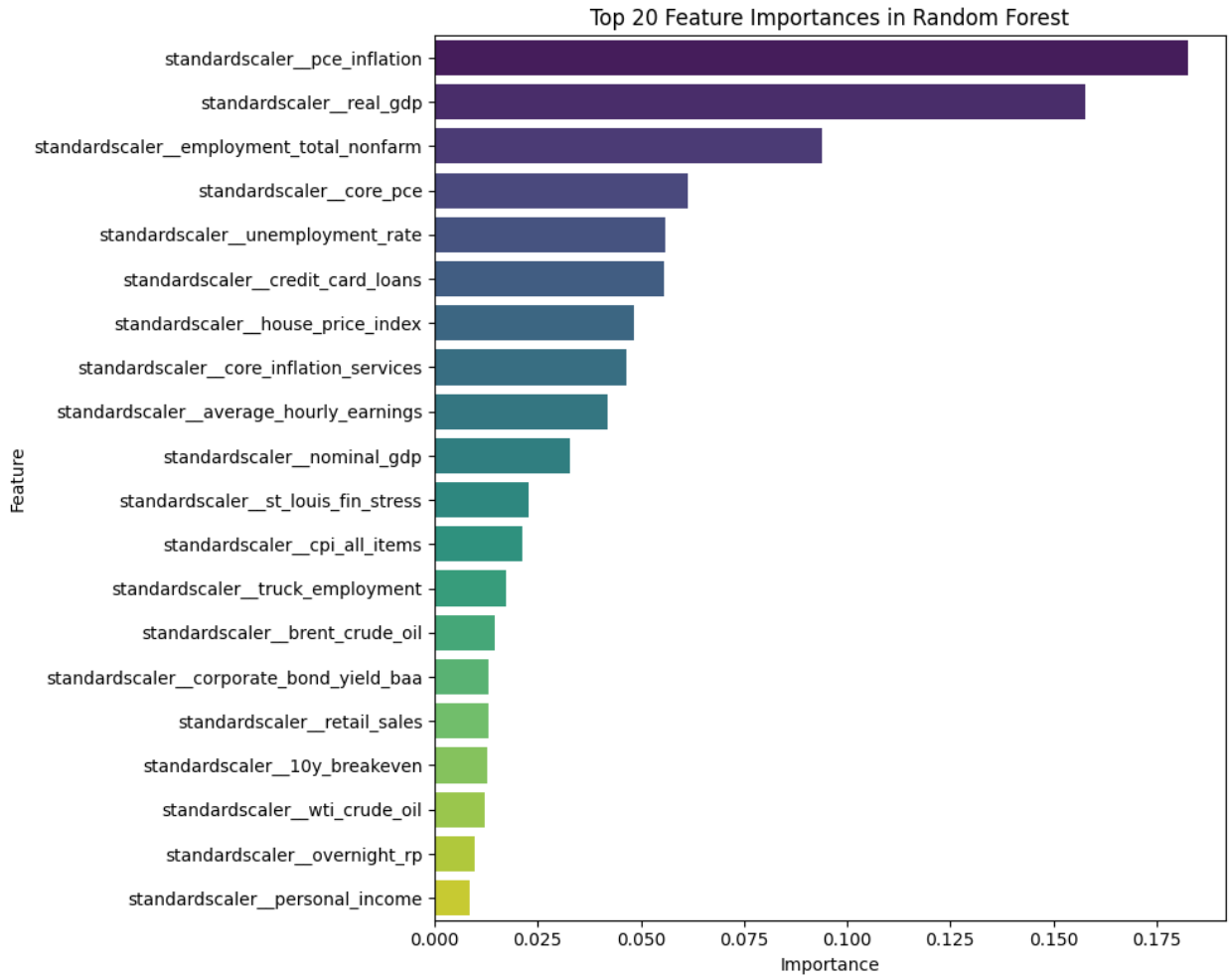
The model achieved exceptionally strong performance, with an R^2 of 1.000 on the test set, a Test MSE of 37.97, and a Test MAE of just 3.83 index points. These results indicate the model captured nearly all the variance in the S&P 500 based on the selected economic fundamentals, showcasing Random Forest's power to model non-linear interactions and variable interactions.

Permutation-based feature importance analysis revealed that the most predictive variables were:

- (pce_inflation) and (core_pce): confirming the role of inflation expectations and core price dynamics.
- (real_gdp) and (employment_total_nonfarm): highlighting the influence of output and labor market fundamentals.
- (unemployment_rate) and (m2_money_stock): reinforcing the importance of cyclical indicators and monetary aggregates.

Notably, seasonal indicators like (quarter_Q3) and firm-level equity data such as (JNJ) or (CVX) had minimal importance in this macro-driven specification, validating the decision to exclude most market-based features.

These results support the conclusion that Random Forest is not only accurate but highly interpretable when paired with proper preprocessing and feature engineering, especially in macro-financial prediction tasks.



Decision Tree Regression: Model Performance and Feature Analysis

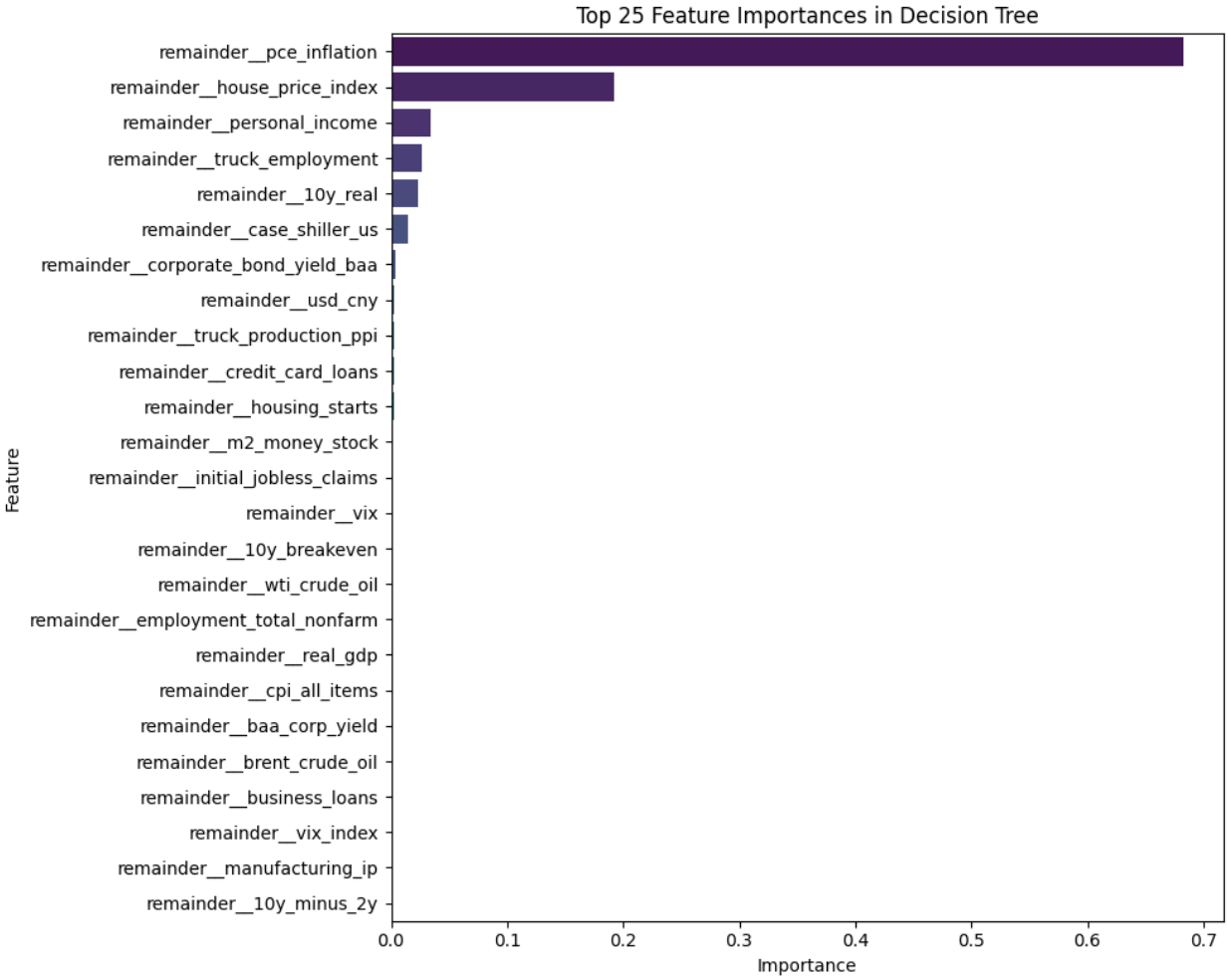
To investigate interpretable non-linear patterns in macroeconomic variables, a Decision Tree Regressor was trained to predict the S&P 500 index. The model was embedded in a preprocessing pipeline that included standardization for all numeric variables and one-hot encoding for the quarterly categorical feature. A grid search across tree depth and node constraints was conducted to optimize model generalization.

The best-performing model achieved a Test MSE of 112.75, Test MAE of 7.80, and a near-perfect R^2 score of 1.000, suggesting strong fit and predictive capability. However, the large gap between training and test MAE (5.05 vs. 7.80) also signals a tendency toward overfitting—an expected tradeoff with decision trees.

Feature importance analysis revealed a highly concentrated structure:

- (pce_inflation) dominated the feature space of total importance, followed by (house_price_index) and (personal_income).
- Other well-known indicators like (employment_total_nonfarm), (truck_employment), and (10y_real) made minor contributions, while many features (e.g., commodity prices, breakeven inflation, and VIX) had near-zero influence.

This skew suggests that the decision tree formed sharp splits around one or two macro variables, potentially at the expense of generalizability. While offering interpretability, this highlights a key limitation of single-tree models: they tend to over-rely on high-variance features, unless carefully regularized or ensembled.



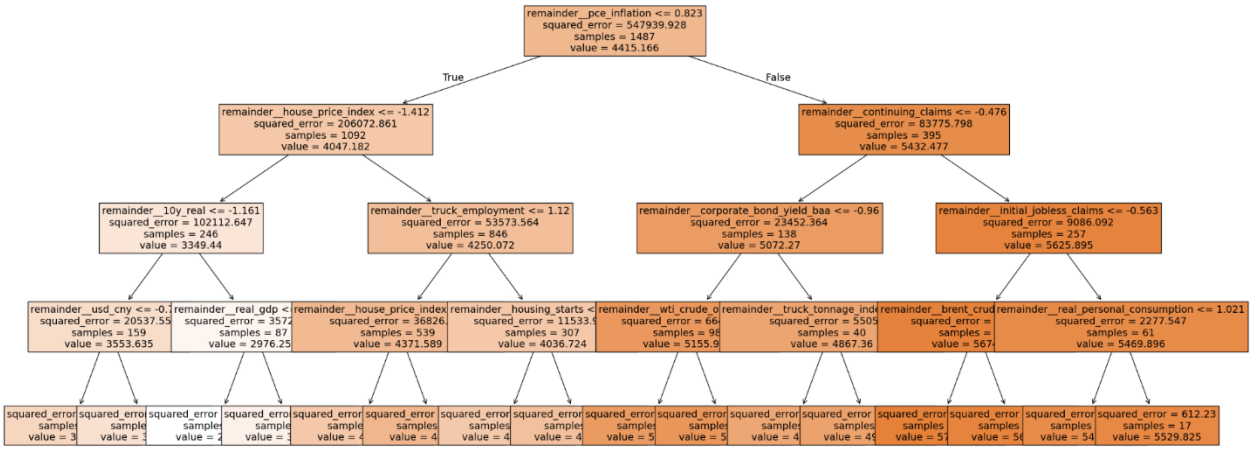
Visualizing Decision Rules in the Decision Tree Model

To enhance model interpretability, a shallow Decision Tree Regressor with a maximum depth of 4 was trained and visualized. The tree structure (Figure X) reveals how macroeconomic variables are hierarchically used to partition the data and predict the S&P 500 index.

At the root node, the model splits on (pce_inflation), affirming its dominant role in explaining stock market variation. This primary decision rule divides the dataset into high- and low-inflation regimes. For observations where inflation is low (≤ 0.823), the next most important factor is the (house_price_index), followed by variables like (10y_real) and (truck_employment)—all of which represent real economic activity and investment-sensitive sectors.

On the opposite branch, where inflation is high, the model splits based on (continuing_claims) and (initial_jobless_claims), suggesting that labor market conditions interact with inflation to shape equity valuations. Further downstream splits incorporate (corporate_bond_yield_baa) and (real_personal_consumption), capturing credit risk and demand-side strength, respectively.

This decision tree provides a transparent sequence of rules that mirrors intuitive macro-financial logic: monetary variables dominate first-level splits, followed by real activity and labor market indicators. Although limited in complexity, this visualization helps bridge the gap between machine learning predictions and traditional economic narrative by explicitly showing how conditions such as “low inflation + strong housing market” or “high inflation + tight labor market” translate into different predicted stock market outcomes.

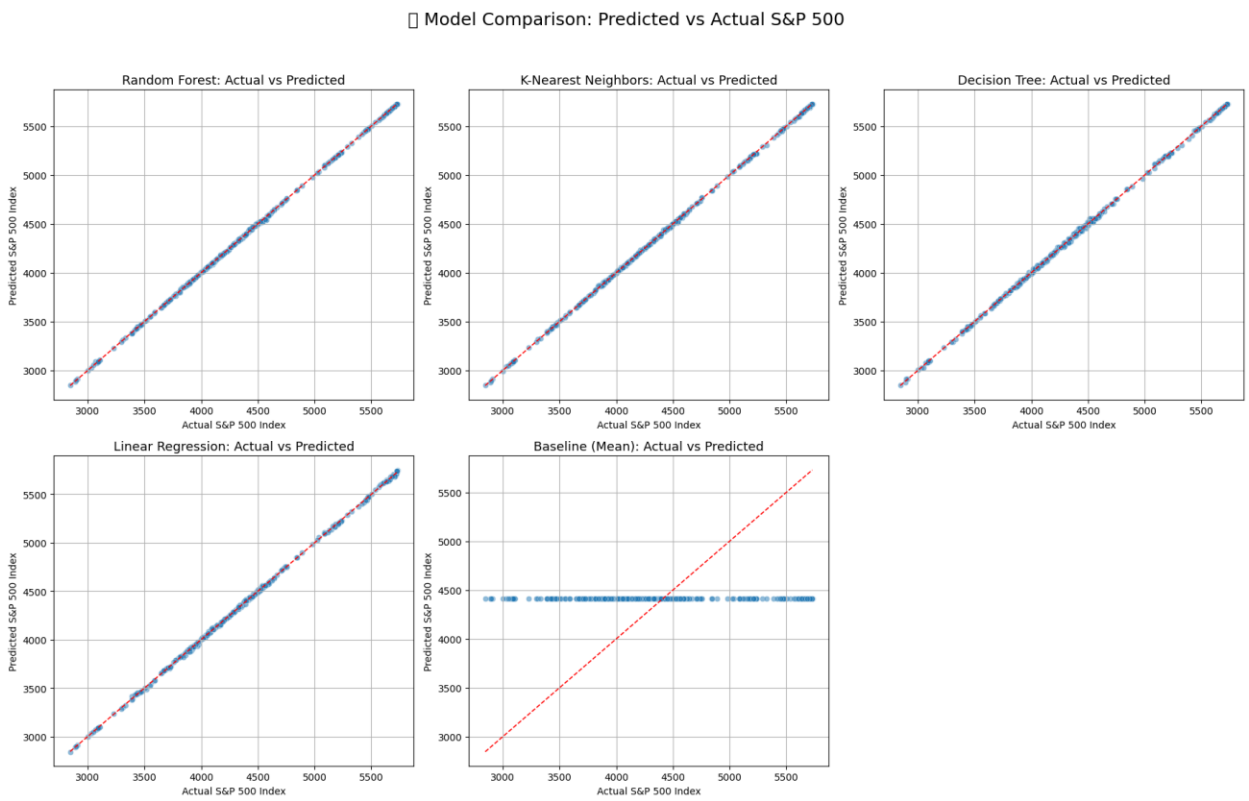


Comparative Visualization: Predicted vs Actual S&P 500 Across Models

To evaluate and compare the predictive performance of all trained models, actual versus predicted values were plotted across the test set (see Figure X). The red dashed line represents the ideal prediction line where actual equals predicted.

- Random Forest, K-Nearest Neighbors, and Linear Regression models all exhibit very tight clustering along the diagonal, indicating high accuracy and minimal bias. This consistency supports their quantitative metrics such as $R^2 \approx 1.000$.
- The Decision Tree model performs well but shows slightly more dispersion than Random Forest, consistent with its higher sensitivity to training variance and lower generalization robustness.
- The Baseline (Mean) model, which predicts a constant mean value, serves as a clear contrast. It fails to capture any variation in the target variable, and the predictions fall on a horizontal line—highlighting the value added by more advanced learning methods.

These visual results reinforce earlier quantitative findings: Random Forest achieved the best overall performance, followed closely by Linear Regression and KNN, while Decision Tree remained interpretable but less stable. The combination of numerical metrics and graphical diagnostics offers a well-rounded validation of model effectiveness.



Cross-Model Comparison: Feature Importance and Learning Dynamics

To compare the internal mechanisms and learning behavior of different predictive models, two types of diagnostics were employed: feature importance analysis and learning curve evaluation.

Feature Importance

Figure X compares the top 15 features identified by Random Forest and Decision Tree models. Both models consistently highlight (pce_inflation) and (real_gdp) as dominant drivers of S&P 500 variation. Random Forest assigns more balanced weights across additional variables such as (employment_total_nonfarm), (unemployment_rate), and (credit_card_loans), capturing broader macroeconomic signals. In contrast, the Decision Tree model is much more concentrated, with (pce_inflation) alone accounting for the majority of its decision structure—consistent with its first-split logic observed in the tree diagram.

K-Nearest Neighbors and Linear Regression are excluded from this comparison, as they do not natively support tree-based feature importance metrics. For KNN, permutation importance was analyzed separately (see Section 5).

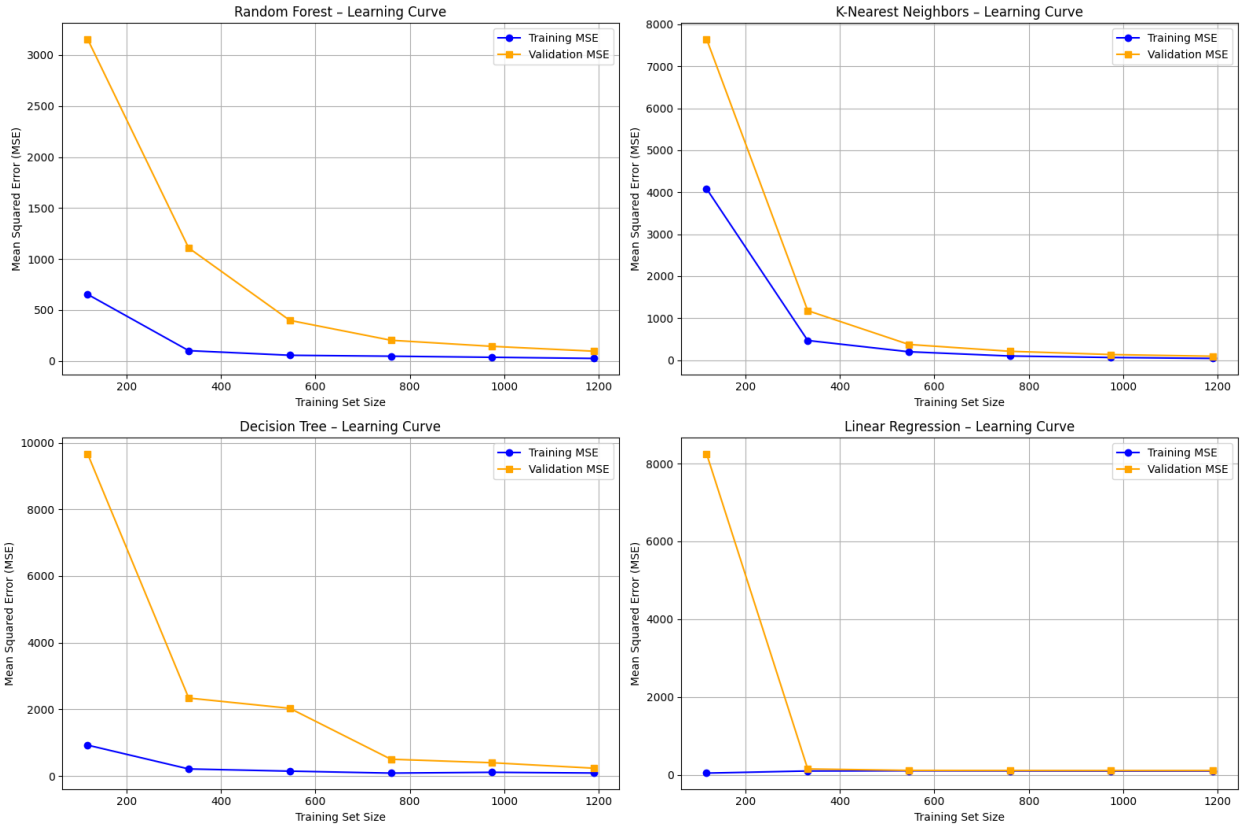
Learning Curves

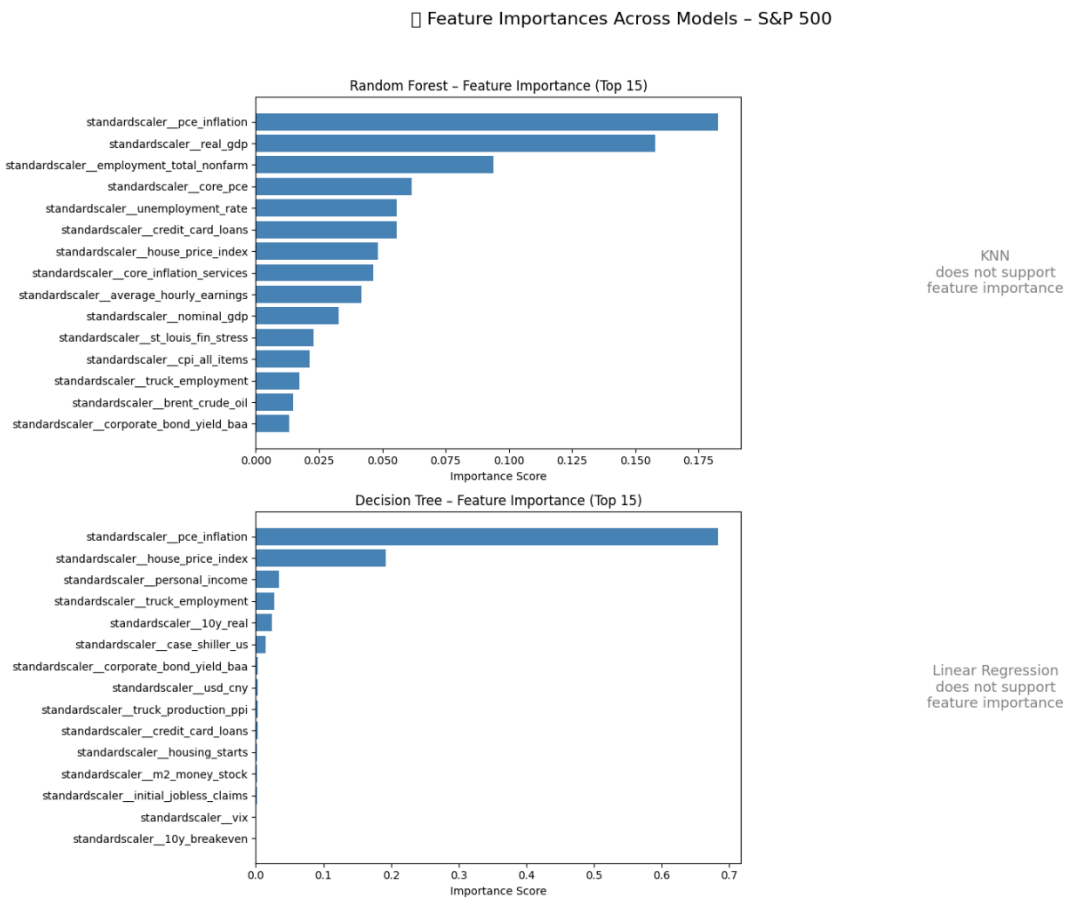
The learning curves in Figure Y reveal key insights about generalization capacity:

- Random Forest achieves low training and validation MSEs with increasing data, and maintains a narrow gap, indicating strong generalization.
- Linear Regression shows excellent stability and converges quickly, though it may underfit slightly due to its linear constraints.
- Decision Tree exhibits significant overfitting on small datasets, which is gradually reduced with more training examples.
- KNN improves with data but retains a higher variance, suggesting sensitivity to local fluctuations in macroeconomic space.

These diagnostics complement previous performance metrics and offer model-specific insights. They confirm that Random Forest not only performs best in accuracy but also balances interpretability and learning efficiency effectively, making it the preferred choice for S&P 500 prediction in this context.

 Learning Curves of Models – Predicting S&P 500



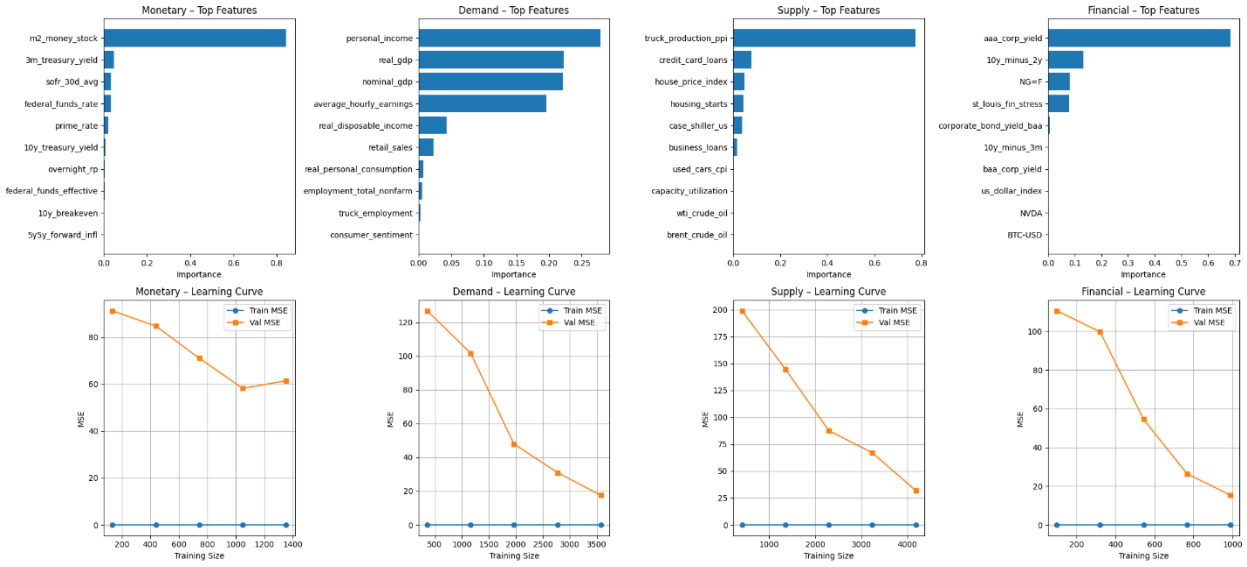


In the prediction task for core PCE, the four feature schemes—monetary variables, demand-side variables, supply-side variables, and financial market & expectation variables—exhibited noticeably different levels of predictive power and structural relevance. The model based on monetary features performed relatively poorly in generalization, with validation errors remaining high despite near-zero training error. Its learning curve flattened early, suggesting limited benefit from additional data. Among these features, (m2_money_stock) dominated in importance, while most interest rate-related indicators contributed little, indicating a skewed internal structure.

By contrast, the demand-side model demonstrated strong generalization capacity, with validation error decreasing significantly as training size increased. Feature importance analysis revealed that (personal_income), (nominal_gdp), and (real_gdp) were the most influential predictors—highlighting the direct relationship between aggregate demand dynamics and core inflation trends.

The supply-side model performed slightly worse than the demand-based model but still showed a reasonable learning trajectory. Notably, (truck_production_ppi) emerged as a dominant feature, possibly capturing cost-push inflation dynamics. The rest of the supply-side variables had only marginal importance.

Lastly, the model based on financial markets and expectations underperformed across all metrics, with high test errors, unstable learning curves, and negative out-of-sample R^2 values. Although features such as (aaa_corp_yield) and (10y_minus_2y) showed some relative importance, financial indicators alone failed to provide a reliable basis for forecasting core PCE movements.



Conclusion

This study developed and evaluated multiple machine learning models to forecast two critical economic indicators: Core PCE inflation and the S&P 500 index. By leveraging a rich macro-financial dataset and carefully partitioned variable schemes, we compared the predictive power and interpretability of linear models, tree-based models, and non-parametric approaches.

The results consistently demonstrate that Random Forest outperformed other models in terms of predictive accuracy and generalization, especially in the S&P 500 forecasting task. Its ability to capture non-linearities and variable interactions, while maintaining interpretability through feature importance analysis, makes it particularly well-suited for macro-financial modeling. Linear Regression, while slightly less accurate, offered valuable economic insights through interpretable coefficients, and K-Nearest Neighbors provided a useful non-parametric benchmark despite sensitivity to noise. Decision Trees, though highly interpretable, were prone to overfitting in small-sample scenarios.

For Core PCE inflation prediction, the demand-side feature scheme exhibited the strongest generalization capacity, suggesting that real economic activity—particularly income and GDP—remains a key driver of underlying inflation trends. Monetary indicators and financial market variables performed relatively poorly, indicating limited standalone explanatory power in this context.

Overall, the comparative model analysis underscores the importance of both methodological rigor and economic domain knowledge in macroeconomic forecasting. Future research may further benefit from hybrid model designs, regularization techniques, and real-time data integration to enhance robustness and responsiveness in dynamic policy and market environments.