Lunamia Bolanos, Pia Jain, Dhwani Agarwal
Professor Koehler
Final Project
December 18, 2025

Predicting Stock Return Direction Using Macroeconomic and Market Indicators

Introduction

This project explores whether macroeconomic indicators and related financial market data can be used to predict the monthly direction of ITB, the iShares U.S. Home Construction ETF. Housing stocks are especially sensitive to changes in interest rates, inflation, labor market conditions, and housing prices, making ITB a useful case study for macro-driven prediction. The goal of the analysis is to determine whether these broader economic signals contain information that helps forecast whether ITB's return will be positive or negative in the following month. Several models with increasing complexity are tested to evaluate whether more flexible, nonlinear approaches improve predictive performance. Overall, linear models perform poorly, while nonlinear classifiers modestly outperform simple baselines in predicting return direction.

Midterm Project

This project builds directly on our midterm analysis, which focused on understanding how macroeconomic forces such as interest rates, inflation, and housing prices shape the U.S. housing market and real-estate equity performance. In the midterm, the emphasis was primarily descriptive as we documented strong co-movement between macro variables and housing-related stocks, while also observing substantial volatility and limited short-term predictability in monthly returns. These findings motivated a shift in strategy for the final project. Rather than continuing with descriptive macro analysis, we moved toward a predictive framework that explicitly tests whether macroeconomic conditions and recent market trends can forecast the direction of stock returns. This transition reflects a deeper engagement with course themes around market efficiency and the limits of predictability, using the midterm results as a foundation to evaluate whether more flexible modeling techniques can extract signals from an otherwise noisy financial environment.

Data Description

The dataset combines monthly macroeconomic indicators with financial market returns to analyze the relationship between economic conditions and housing equity performance. Although the raw data extend back to 2015, the analysis focuses on the period beginning in 2020 to ensure consistency across all macroeconomic, housing, and financial series and to reflect the more recent economic environment. All variables are aligned at a monthly frequency using month-end observations, allowing the construction of a unified time-series dataset that is appropriate for predictive modeling and reduces issues related to timing mismatches across data sources.

Macroeconomic data are sourced from the Federal Reserve Economic Data (FRED) database and are chosen to capture key dimensions of the economic environment that influence housing and real-estate equities. Inflation is measured using the Consumer Price Index (CPI), labor market conditions are represented by the unemployment rate, and monetary policy is proxied by the effective federal funds rate. For each of these variables, both the level and the month-to-month change are included in the dataset. This structure allows the models to distinguish between persistent economic conditions and short-term macroeconomic shocks, which may have different effects on asset prices.

Housing market conditions are incorporated using the Federal Housing Finance Agency (FHFA) House Price Index. The index is aggregated to a monthly frequency and converted into percentage changes to emphasize housing price momentum rather than long-term price levels. Financial market data are obtained from Yahoo Finance and include monthly returns for ITB, the primary asset of interest, along with returns for several related assets that capture broader sectoral and macroeconomic exposure. These include industrial real estate, data center real estate, large banks, and energy, which are economically linked to housing activity through financing conditions, infrastructure demand, and input costs. In addition to raw monthly returns, three-month rolling average returns are calculated for each asset to capture short-term momentum and reduce idiosyncratic volatility.

The modeling targets are defined to support both regression and classification tasks. For regression, the target variable is ITB's next-month return, which allows the models to estimate the magnitude of future performance. For classification, a binary indicator is constructed that equals one if ITB's next-month return is positive and zero otherwise. This dual approach allows for a direct comparison between predicting return magnitude and forecasting return direction, with the latter often being more relevant in applied investment and risk management settings.

Models and Methods

To avoid look-ahead bias, the data are split using a time-based approach in which the final twelve months are reserved as a holdout test set, and all earlier observations are used for model training. This structure reflects a realistic forecasting environment in which only past information is available when making predictions about future returns. Before modeling, all predictor variables are standardized to ensure that differences in scale do not disproportionately influence model estimation, particularly for distance-based methods.

The analysis begins with linear regression to predict ITB's next-month return and logistic regression to classify whether the return is positive or negative. These models serve as baseline benchmarks and impose linear relationships between macroeconomic variables, market signals, and ITB performance. To account for potential nonlinearities and local patterns in the data, more flexible machine learning models are then introduced. K-Nearest Neighbors (KNN) classification is used to capture similarity-based relationships across months with comparable macro-financial conditions, while a Random Forest classifier is applied to model nonlinear interactions and feature combinations through an ensemble of decision trees. Model performance is evaluated using regression metrics such as R2 and mean absolute error, as well as classification accuracy and confusion matrices, which provide insight into each model's ability to correctly identify both upward and downward return periods.

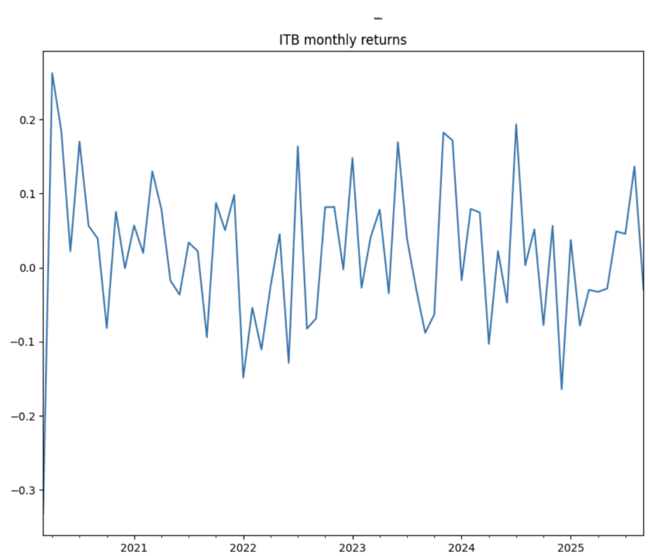

Figure X plots monthly returns for ITB over the sample period and highlights the high volatility and noise in short-horizon housing equity performance. Returns frequently swing between positive and negative values, with several large spikes and drawdowns, particularly during periods of macroeconomic uncertainty.

...

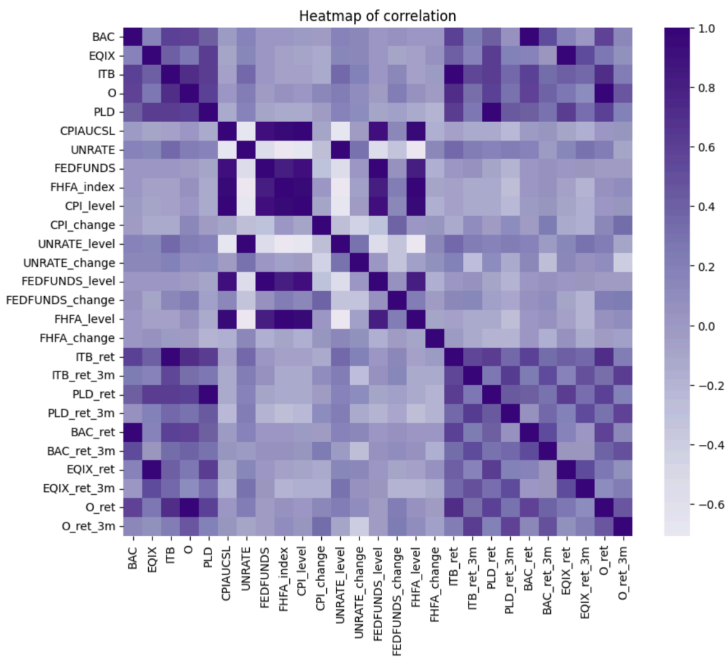

Figure X presents a correlation heatmap of macroeconomic variables, housing indicators, and asset returns. The heatmap shows strong co-movement among related real-estate and market assets, as well as expected relationships among macroeconomic variables.

Results and Interpretation

The linear regression model performs poorly, producing a negative R2 value, which indicates that it predicts next-month returns worse than simply using the historical average. While the mean absolute error is reasonable given the inherent volatility of ITB returns, the model consistently fails to capture large positive and negative movements. This suggests that the relationship between macroeconomic variables and short-horizon equity returns is not well described by a linear framework. Logistic regression similarly performs poorly, achieving the same accuracy as a baseline model that always predicts the most common outcome. Examination of the confusion matrix shows that the logistic model predicts only positive months, resulting in perfect recall for upward movements but no ability to correctly identify down months. This behavior reflects the limitations of linear decision boundaries when applied to noisy and highly volatile financial data, particularly with a small sample size.

Nonlinear models produce more promising, though still limited, results. The K-Nearest Neighbors (KNN) classifier improves directional accuracy to approximately 60-67 percent. Unlike logistic regression, KNN predicts both up and down months and avoids false negatives in the test set, indicating that localized similarities in macroeconomic and market conditions contain useful predictive information. The Random Forest classifier achieves comparable accuracy, correctly classifying roughly two-thirds of observations and providing a more balanced distribution of predictions across classes. However, because the test set contains only twelve observations, the Random Forest does not meaningfully outperform KNN and may be vulnerable to overfitting. To test this conclusion, we attempted to fine-tune the KNN model by testing different k numbers from 1 to 11 and received the best accuracy with k=5. As mentioned, we did not aim to overfit, so we finalized our conclusion that while nonlinear models can extract modest signals from the data, predictive performance remains constrained by noise and limited sample size.

Conclusion

This analysis shows that macroeconomic and cross-asset financial data contain limited but meaningful information for predicting the direction of housing equity returns. Linear models are largely ineffective in this setting, reinforcing the idea that short-horizon stock movements are difficult to explain using simple linear relationships. In contrast, nonlinear approaches such as KNN and Random Forests provide modest but consistent improvements over weak baselines, suggesting that localized patterns and nonlinear interactions can capture some predictive signal. Despite these gains, overall performance remains constrained by the small sample size and the inherent noise in monthly equity returns, which limits the reliability of any single model's predictions. Future work could expand the dataset, incorporate additional housing-specific variables such as mortgage rates, housing starts, or building permits, and apply rolling or cross-validation techniques to better evaluate model stability over time. Taken together, the project highlights both the potential and the practical limits of machine learning in financial prediction, aligning with broader course themes around market efficiency and the challenges of extracting signals from noisy economic data.