Do Behavioral Narratives Boost Forecast Accuracy in Machine Learning-Based Stock Models?

The thesis written under the supervision of
dr hab. Tomasz Kopczewski
mgr Ewa Weychert
Faculty of Economic Sciences

Marco Ayuob 465767
Data Science and Business Analytics

Motivation

- **Markets are not only related to hard indicators**— they are also related to stories, and attention.
- **Efficient Market Hypothesis (EMH) Fama (1970)–** Stock prices always include all information
- **Narrative Economics Shiller (2019)** – argues that stories can explain why markets sometimes move without clear economic fundamentals. Tools like Google Trends can help track the spread and influence of these narratives.

Research Questions

- Can “market gossip” boost forecast accuracy represented by Google Trends?
- How does the uncertainty period (highly volatile) affect the behavioral prediction accuracy?
- We investigate which features most contribute to the predictions using SHAP.

Data & Method

Source Data	Behavioral Data	Source: Google Trends via TrendEcon and gtrendsR (GitHub, API-based queries)
	Macroeconomic Indicators	1. U.S. Census Bureau (Unemployment, Net Trade, Oil Prices) 2. U.S. Bureau of Labor Statistics (CPI) 3. Federal Reserve Bank of St. Louis (T-Bill Rate, Exchange Rate, GEPU Index)
	Stock Data	Monthly closing prices of 10 major U.S. companies (2013–2024), quantmod package in R
Method	Feature Engineering	Lag features, rolling stats, and behavioral interactions (e.g., Google × Oil), motivated by:(Beyaz et al. (2018)
	ML Algorithms	Random Forest, ElasticNet, XGBoost, motivated by: Wu et al. (2023), Patel et al. (2015), and Cheng (2023) -----Gu, Kelly, & Xiu, 2020
	Methods of tuning and cross-validation	Grid search – 10-fold CV
	Model Interpretation	SHAP analysis for feature impact and explainability

Results

- **Q1 Can “market gossip” boost forecast accuracy represented by Google Trends?**

  Yes, it was highlighted among the top features based on XGBoost Tree Importance (Gain-Based), which is the best model after many assessments.

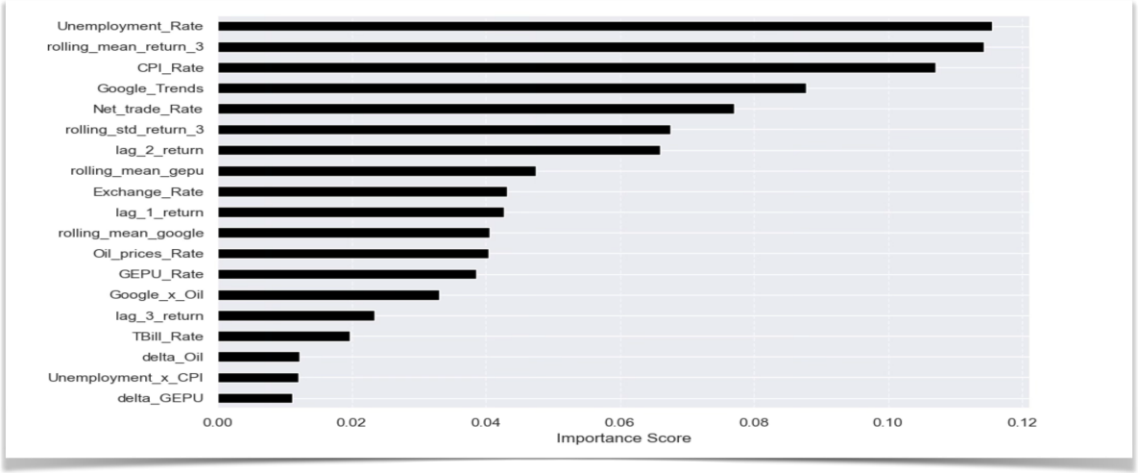
- **Q2 How does the uncertainty period (highly volatile) affect the behavioral prediction accuracy?**

  During the high uncertainty times, stock prices become more volatile, which makes the investor rely more on behavioral factors than on the numbers (traditional indicators), according to SHAP results

- **Q3 We investigate which features most contribute to the predictions using SHAP.**

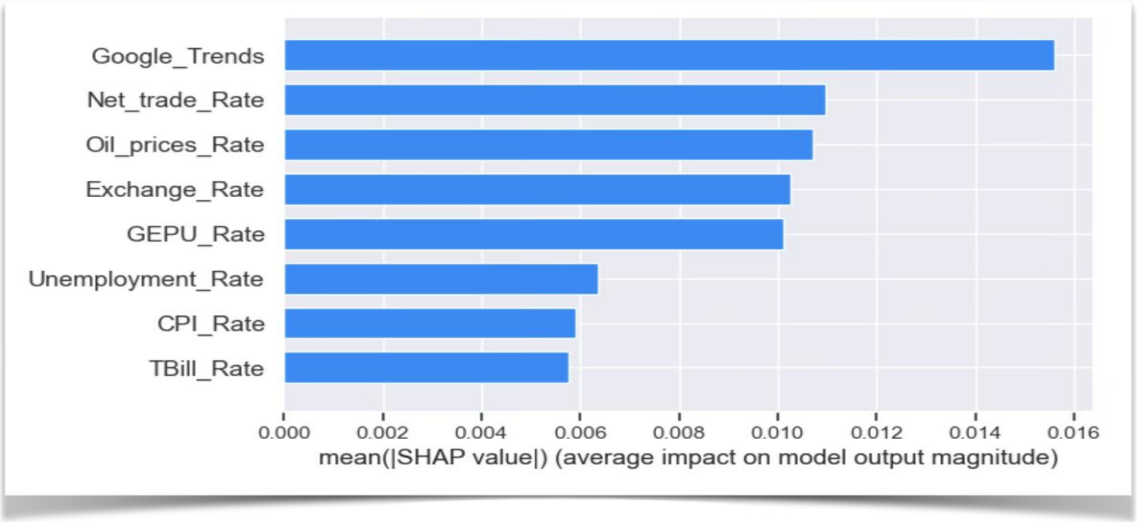
  SHAP shows Google Trends ranked top 1 based on SHAP-Based Feature Contribution (Model-Agnostic). It confirms their strong role in prediction accuracy.

Results



XGBoost Tree Importance (Gain-Based)

- Model-specific (only tree-based like XGBoost)
- Based on how often and how well a feature splits the data in trees



SHAP-Based Feature Contribution (Model-Agnostic)

- Model-agnostic (works with any model, linear or nonlinear)
- Based on game theory – shows each feature's contribution to each prediction

Thank you