Predicting S&P 500 Stocks Return with Firm Characteristics and Macroeconomics Data

Name: Yifan Lu
Institute: DSI
Date:12/5/2023
GitHub Repo: https://github.com/HelenLumi/DATA1030-Project

Recap: Problem Statement

- Predicting S&P 500 stocks return using firm characteristics and macroeconomic data
    - Regression Problem
- **Quantitative equity strategies**
    - Multi-factor model from fundamental, macroeconomic and technical data

$$\mathrm{E}_t(r_{i,t+1}) = g^{\star}(z_{i,t})$$

Return of stock i on time t+1

P-dimensional predictor variables for stock i on time t

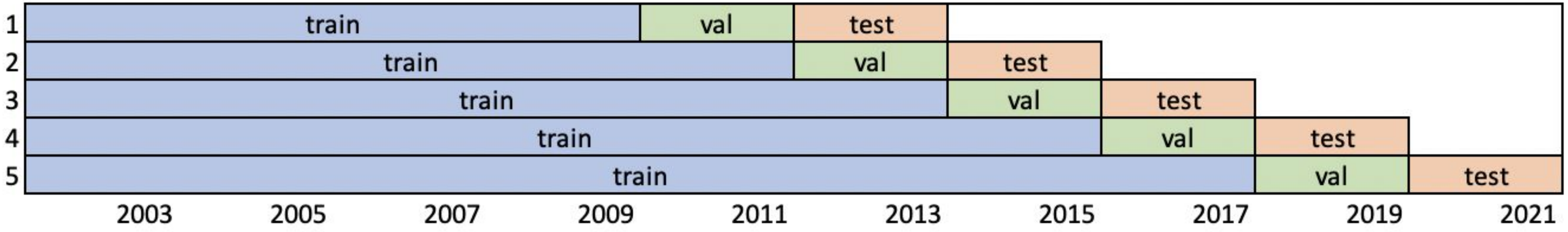
- Important for investors and portfolio managers to manage financial risk, hedge against market drawdown, etc.

Recap: Data

- Data
  - Stock return data: 500 stocks each month, from Mar. 2003 to Dec. 2021
  - 94 Firm Characteristics(FC) and 10 Macroeconomic variables
  - Shape: (80,211, 97)
- Preprocessing
  - **Cross-sectional Median** for missing data
  - **Rank-Normalization** - FC
  - **One-hot Encoder** - Industry variable ‘sic2’
- EDA
  - Lagged 5-month return exists autocorrelation
  - Exclude 10 strongly correlated features
  - Outliers and ambiguous relationship between feature and target
- Preprocessed df: (80,211, 144)

Cross Validation

- **5-Fold Recursive Evaluation Scheme** for time-series data



➔ Recursively increase the training sample to retain the entire history.

➔ Maintain a fixed-size/rolling validation and test sample.

ML models

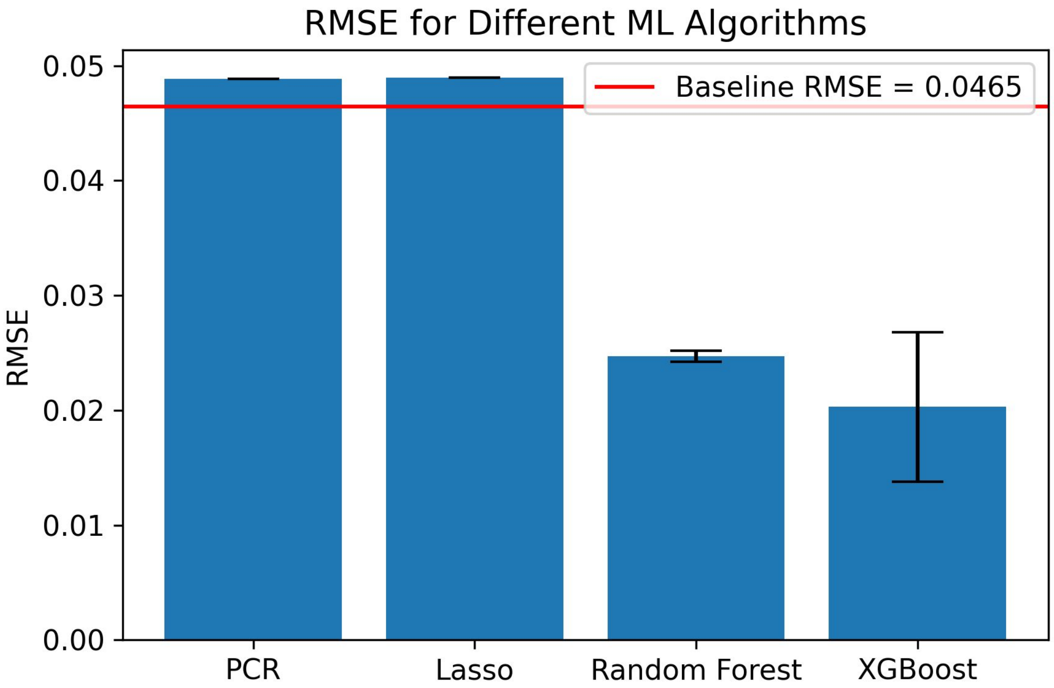
- Hyperparameters

[random_state=42*i]

Principal Components Regression (PCR)	PCA(): ‘n_components’: [1,3,5,7] ‘svd_solver’: 'randomized'
LR with L1 Regularization (Lasso)	'alpha': np.logspace(-4,-1,10)
Random Forests Regression	'n_estimators': [200], 'max_depth': [1, 2, 3, 6], 'max_features': [3, 5, 10, 20]
XGBoost	‘n_estimators’: [10000], ‘early_stopping_rounds’: [50], 'max_depth': [1, 2, 3], 'learning_rate': [.01, 0.1], 'subsample':[0.66]

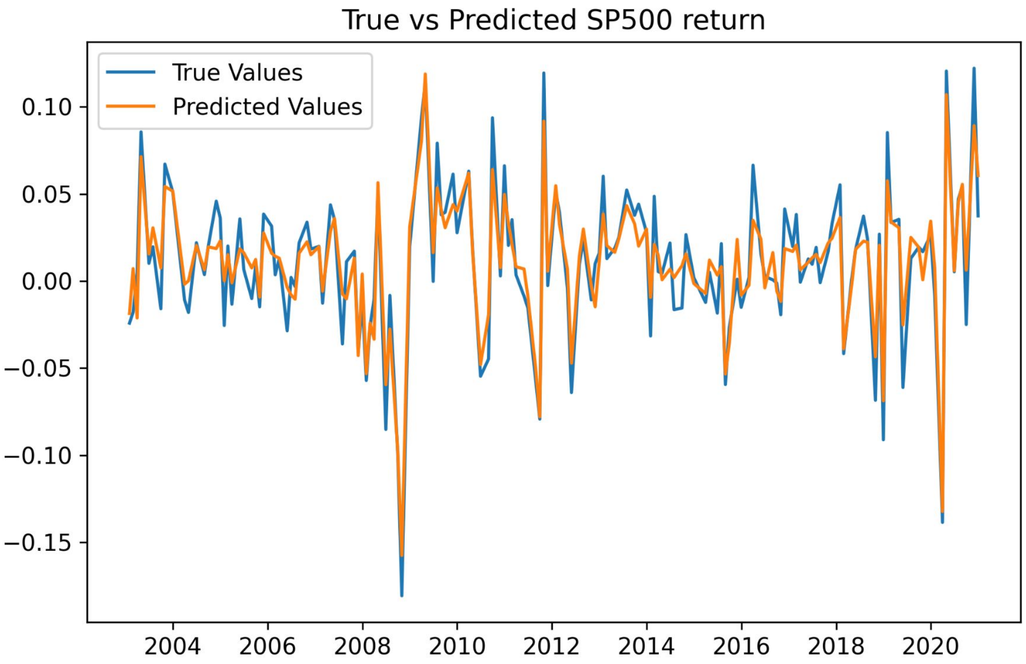
Results

- Baseline: dummy_regressor(strategy='mean')



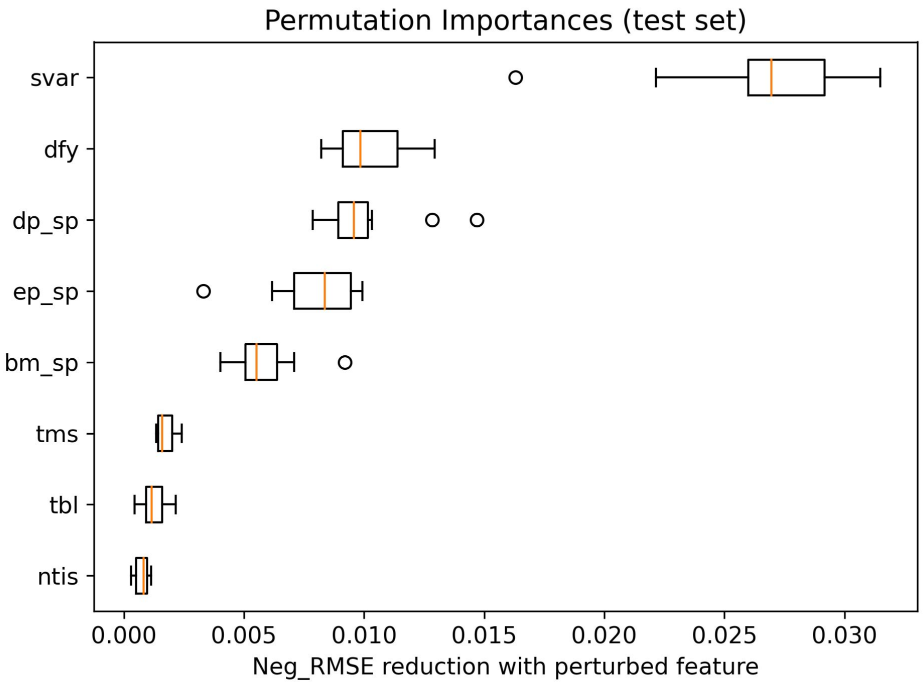
Results

- Plot of True vs Predicted Value - XGBoost Regressor

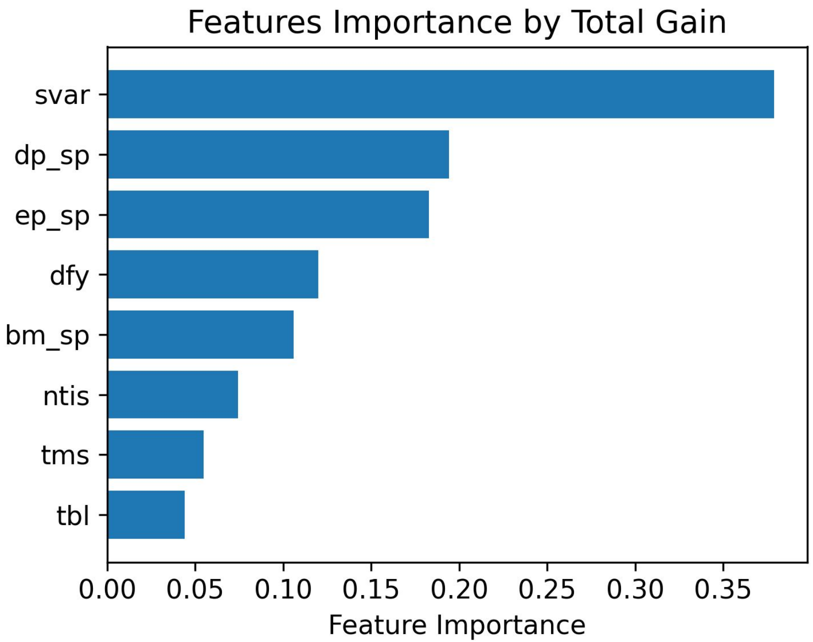


Interpretability

- Permutation Importances



- XGBoost - Total Gain



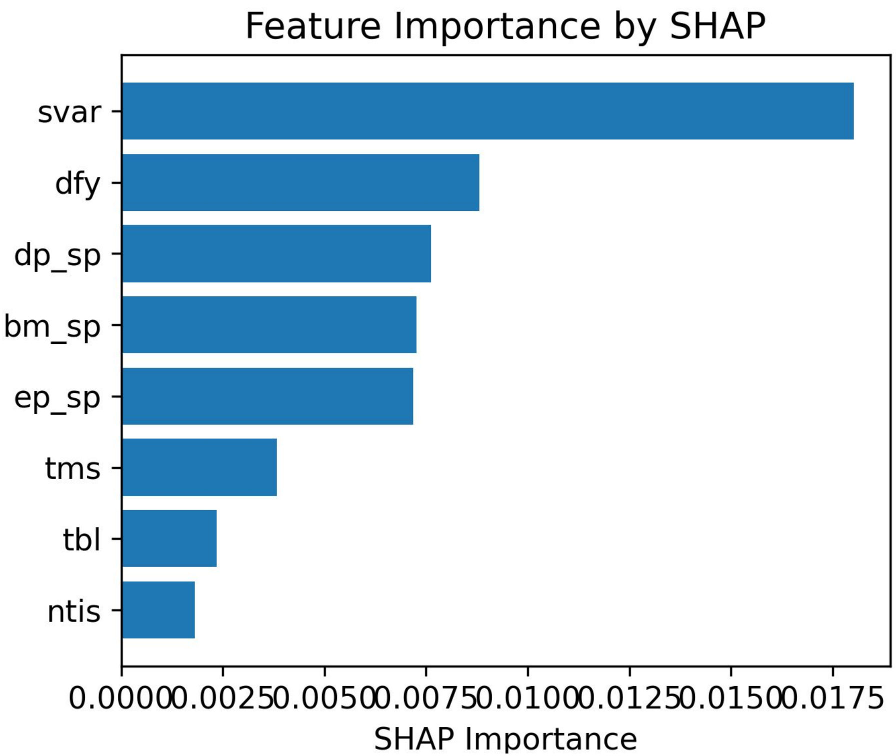
Interpretability

- SHAP - global importance

Top 3 Important features

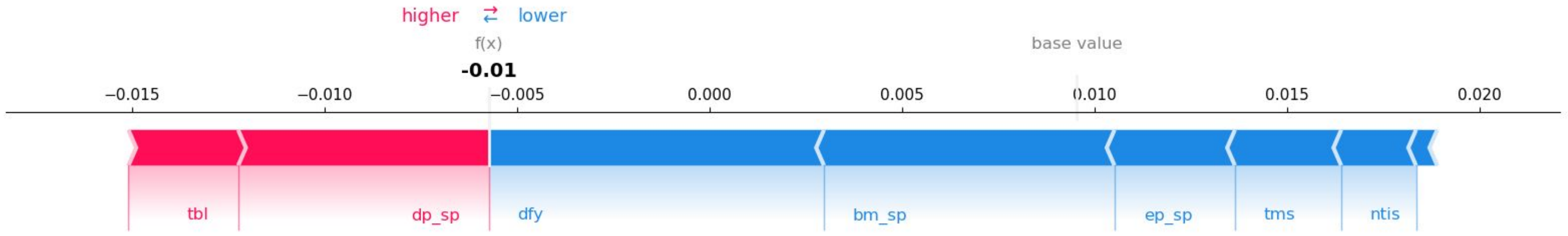
1. svar: stock variance
2. dp_sp: dividend-price ratio
3. dfy: default spread

→ macroeconomic variables



Interpretability

- SHAP - local importance



→ dfy typically drives return lower, dp_sp higher

Outlook

- Feature Engineering - Feature interaction
    - E.g. Product of firm characteristics and macroeconomics data
- Include more lagged return - better capture the trend in historical data
    - E.g. Fit a two-stage model that uses ARIMA and XGBoost
- Try longer time horizon to test the model robustness.

Thank you!

Q&A