INTEGRATING SENTIMENT ANALYSIS WITH STOCK DATA FOR PREDICTIVE MODELING

Chan Tony Yuen Yeung

INTRODUCTION

Why Sentiment Analysis

As a Market Indicators to provide a quantitative measure of market sentiment to enhance the **potential profit opportunity** for **short-frequent trading**.

Objectives

- Use **sentiment analysis** as one of the features for **price movement prediction**.
- Test with different machine learning models to predict the overall movement of the next rebalance window.
- Adjust the portfolio according to the predictions (i.e. True Positive and True Negative)

Methodology

- **Feature Selections:** Test with two feature sets
- Portfolio: Design an MVO portfolio as our main portfolio
- **Classification:** Use classification to predict the upward/downward movement of the stock price. Then, find the most frequent label as the label of the rebalance period.
- **Testing:** Set the rebalance period as 10 days and the lookback period as 50 days (for MVO portfolio). Use back testing to test the performance in 23 windows (230 days).

SENTIMENT ANALYSIS ON STOCK-RELATED TWEETS

In this algorithm, we use the retrieved tweets dataset to do the analysis.

1. Filter out the stocks with fewer than 500 tweets
2. Iterates through each tweet, normalizes the text, and calculates sentiment scores (We use the **SentimentIntensityAnalyzer** from the nltk.sentiment.vader)

STOCK TREND PREDICTION AND FEATURE ENGINEERING

Feature Set 1:
Stock Data, Sentiment Score, Log Return, Moving Average (7 and 14 days)

	LR	RF	AB	LR_PCA	RF_PCA	AB_PCA
AAPL	0.391304	0.565217	0.521739	0.391304	0.608696	0.608696
AMD	0.565217	0.217391	0.217391	0.782609	0.217391	0.782609
AMZN	0.391304	0.565217	0.608696	0.521739	0.521739	0.521739
GOOG	0.608696	0.347826	0.304348	0.652174	0.652174	0.347826
META	0.565217	0.434783	0.347826	0.565217	0.565217	0.565217
MSFT	0.478261	0.391304	0.347826	0.391304	0.391304	0.391304
NIO	0.434783	0.521739	0.739130	0.521739	0.521739	0.521739
PG	0.391304	0.478261	0.434783	0.391304	0.608696	0.608696
TSLA	0.608696	0.478261	0.521739	0.478261	0.478261	0.478261
TSM	0.565217	0.608696	0.434783	0.304348	0.304348	0.695652
Model Mean	0.500000	0.460870	0.447826	0.500000	0.486957	0.552174

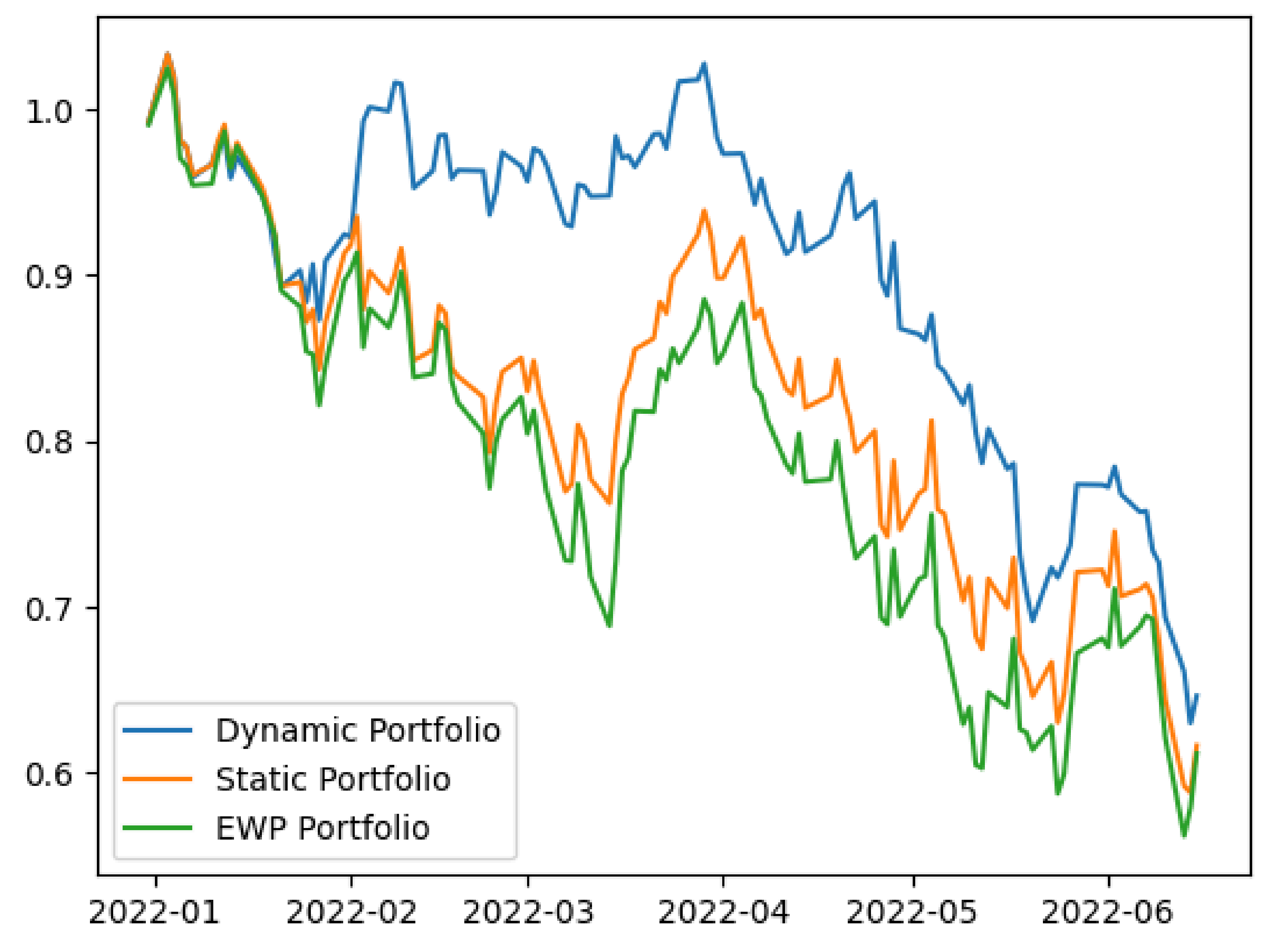
Feature Set 2:
Feature Set 1, MACD, 14SD, Upper/Lower Band, Log Momentum

	LR	RF	AB	LR_PCA	RF_PCA	AB_PCA
AAPL	0.391304	0.608696	0.608696	0.391304	0.391304	0.608696
AMD	0.565217	0.217391	0.347826	0.782609	0.782609	0.782609
AMZN	0.391304	0.608696	0.695652	0.521739	0.521739	0.521739
GOOG	0.608696	0.347826	0.347826	0.652174	0.565217	0.565217
META	0.565217	0.434783	0.434783	0.565217	0.565217	0.565217
MSFT	0.478261	0.478261	0.521739	0.391304	0.391304	0.391304
NIO	0.434783	0.652174	0.608696	0.521739	0.521739	0.521739
PG	0.391304	0.434783	0.478261	0.391304	0.608696	0.608696
TSLA	0.608696	0.478261	0.608696	0.478261	0.521739	0.478261
TSM	0.565217	0.478261	0.391304	0.304348	0.304348	0.304348
Model Mean	0.500000	0.473913	0.504348	0.500000	0.517391	0.534783

PORTFOLIO OPTIMIZATION

- Dynamically optimize the protfolio weight using MVP using the previous prediction
- Compare with the static and EWP portfolios

	Dynamic Portfolio	Static Portfolio	EWP Portfolio
Annual Return	-0.5637	-0.6033	-0.6112
Annual Volatility	0.3263	0.4061	0.4296
Sharpe Ratio	-2.3753	-2.0702	-1.9814
Cumulative Return	0.6463	0.6163	0.6116
Max Drawdown	-0.3474	-0.3823	-0.3958



MVO Portfolio Design

Objective Function:

$$\max \quad w \cdot u - 0.1\sigma^2$$

Constraints:

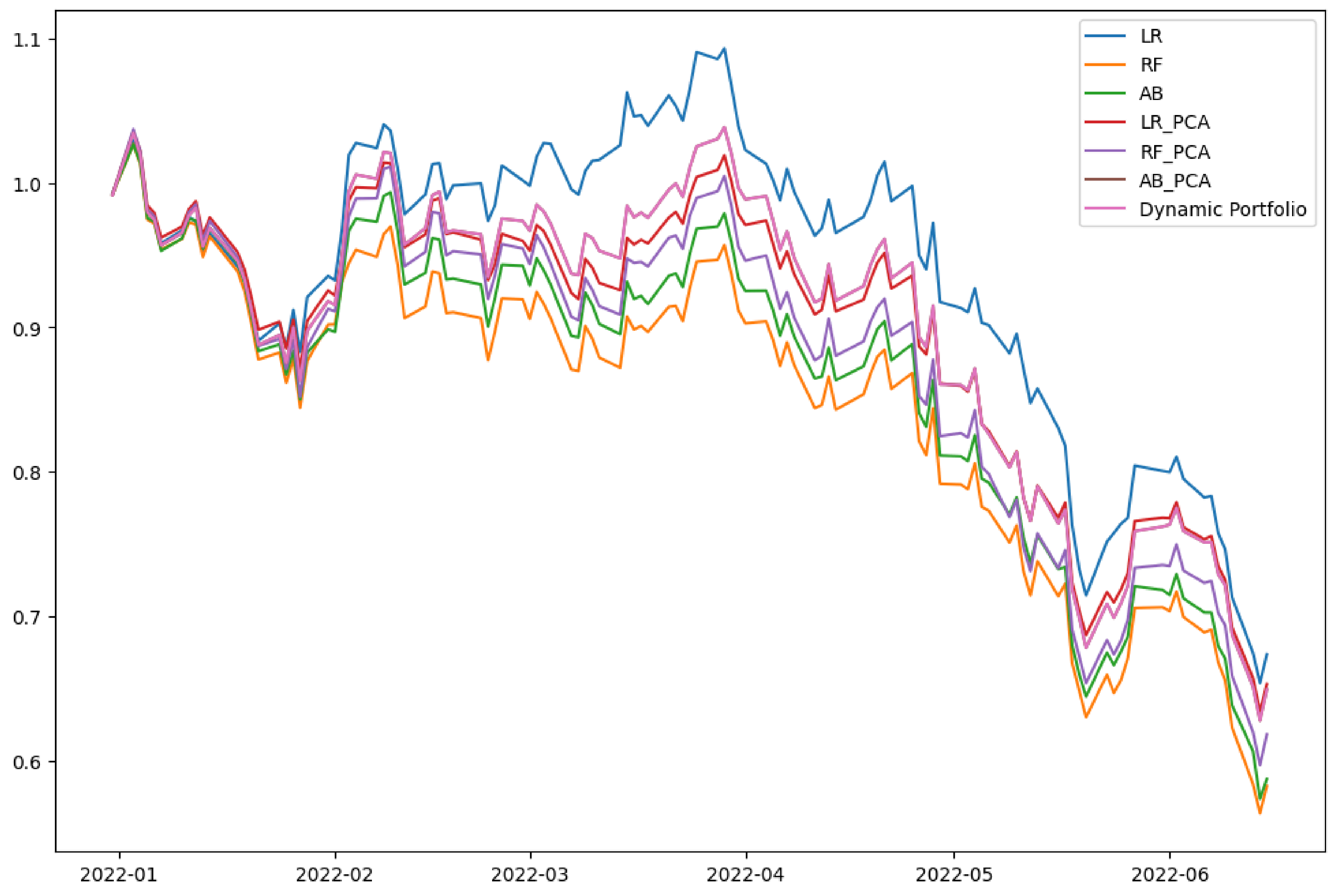
$$\begin{aligned} s.t. \quad & w \geq -0.2, \\ & w \leq 0.2, \\ & \sum w = 1, \\ & \sigma^2 \leq 0.1^2, \\ & \frac{|w_t - w_{t-1}|}{2} \leq 0.15 \end{aligned}$$

ADJUSTING PORTFOLIO WEIGHTS BASED ON PREDICTIONS

Match_weight = 1.2

Weighting changes afterapplying Random Forest

Cumulative Return

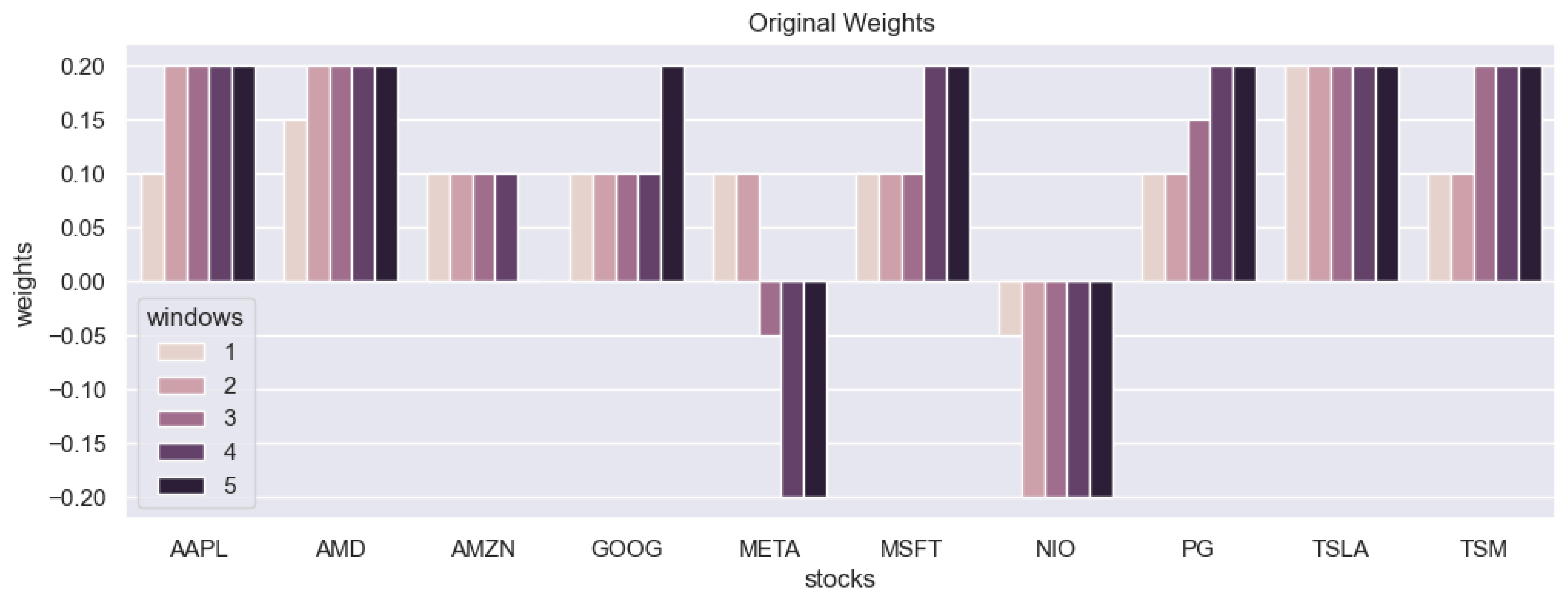


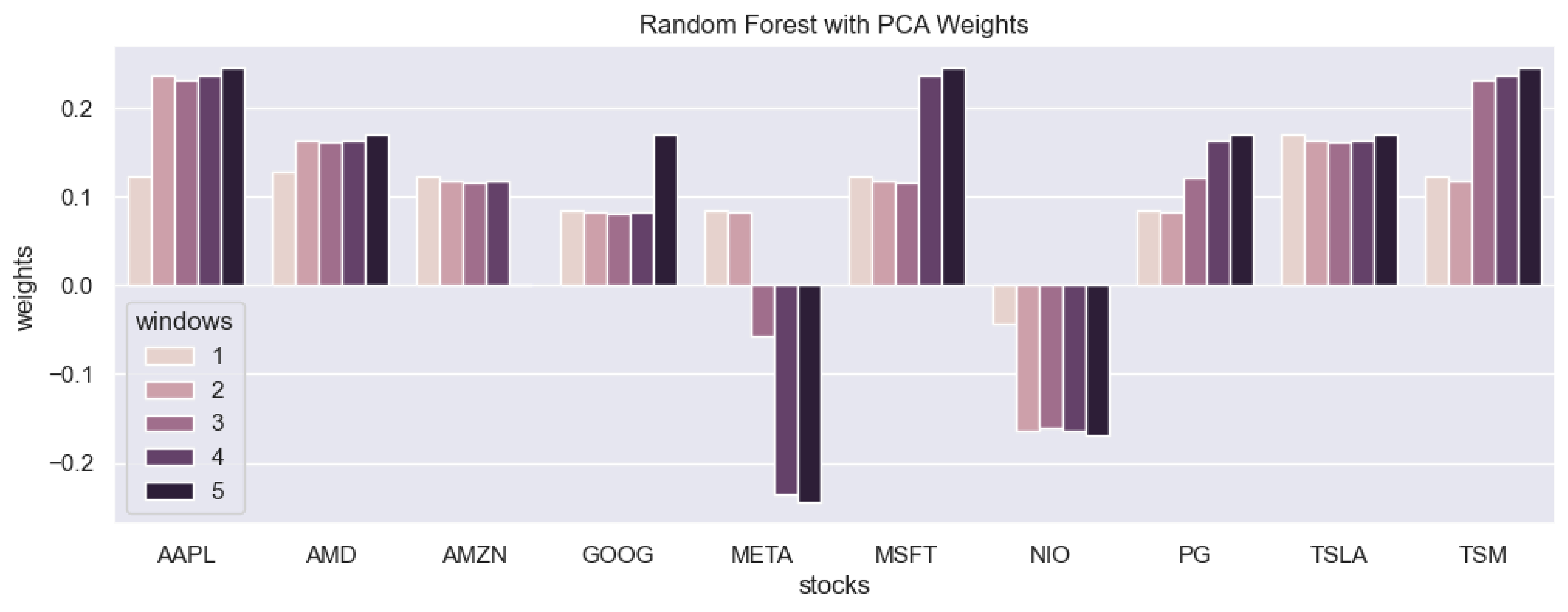
Performance Metric

	LR	RF	AB	LR_PCA	RF_PCA	AB_PCA
Annual Return	-0.5377	-0.6204	-0.6164	-0.5564	-0.5925	-0.5626
Annual Volatility	0.3342	0.3236	0.3244	0.3217	0.3462	0.339
Sharpe Ratio	-2.139	-2.8268	-2.7867	-2.3629	-2.4156	-2.2658
Cumulative Return	0.6738	0.5826	0.5875	0.6532	0.6183	0.6494
Max Drawdown	-0.3643	-0.3855	-0.3794	-0.3458	-0.3731	-0.351

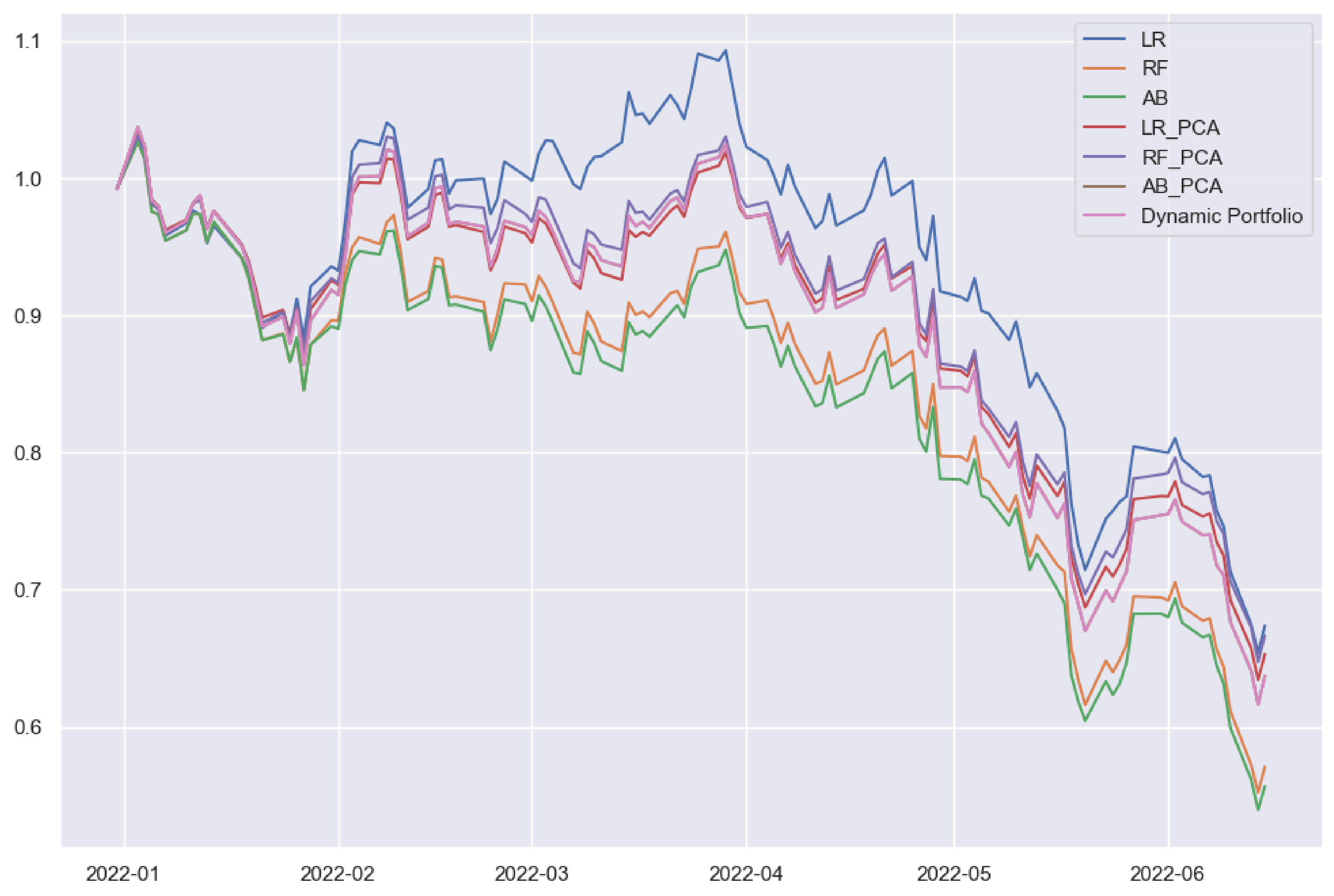
BACKTESTING ON FEATURE SET 2 USING RANDOM FOREST

Weighting changes afterapplying Random Forest





Cumulative Return



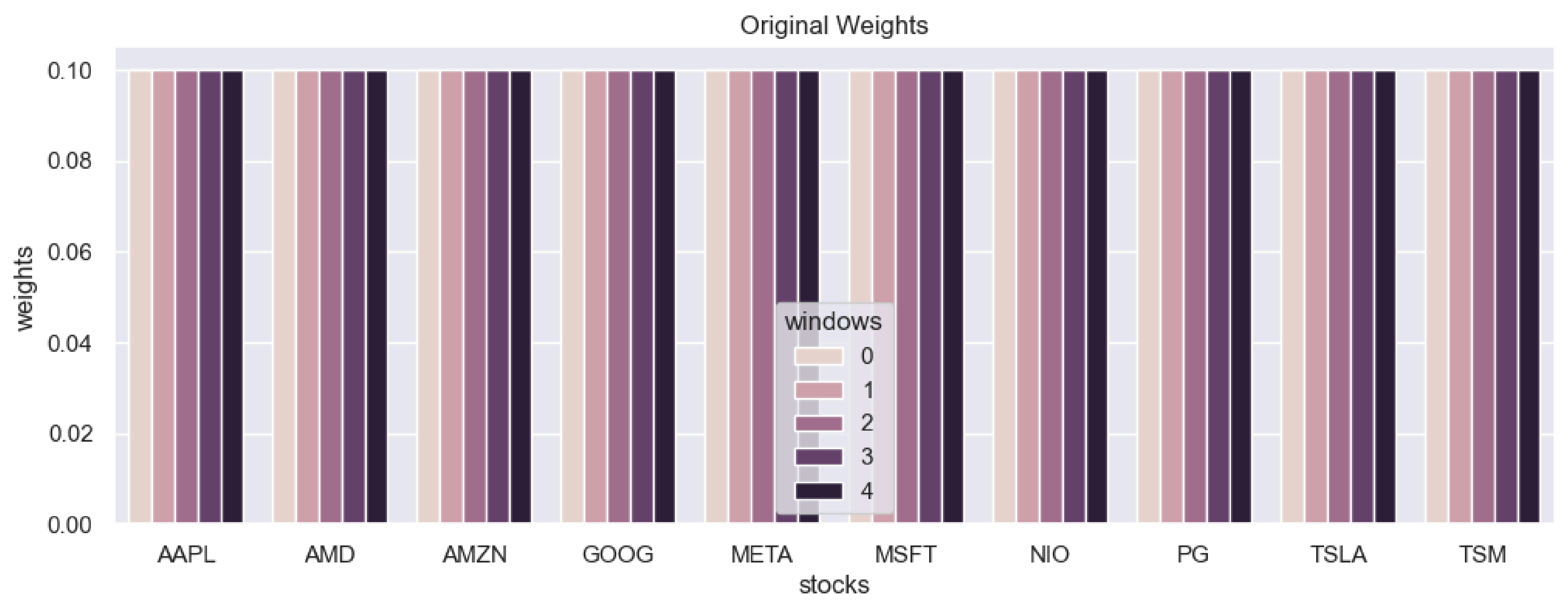
Performance Metric

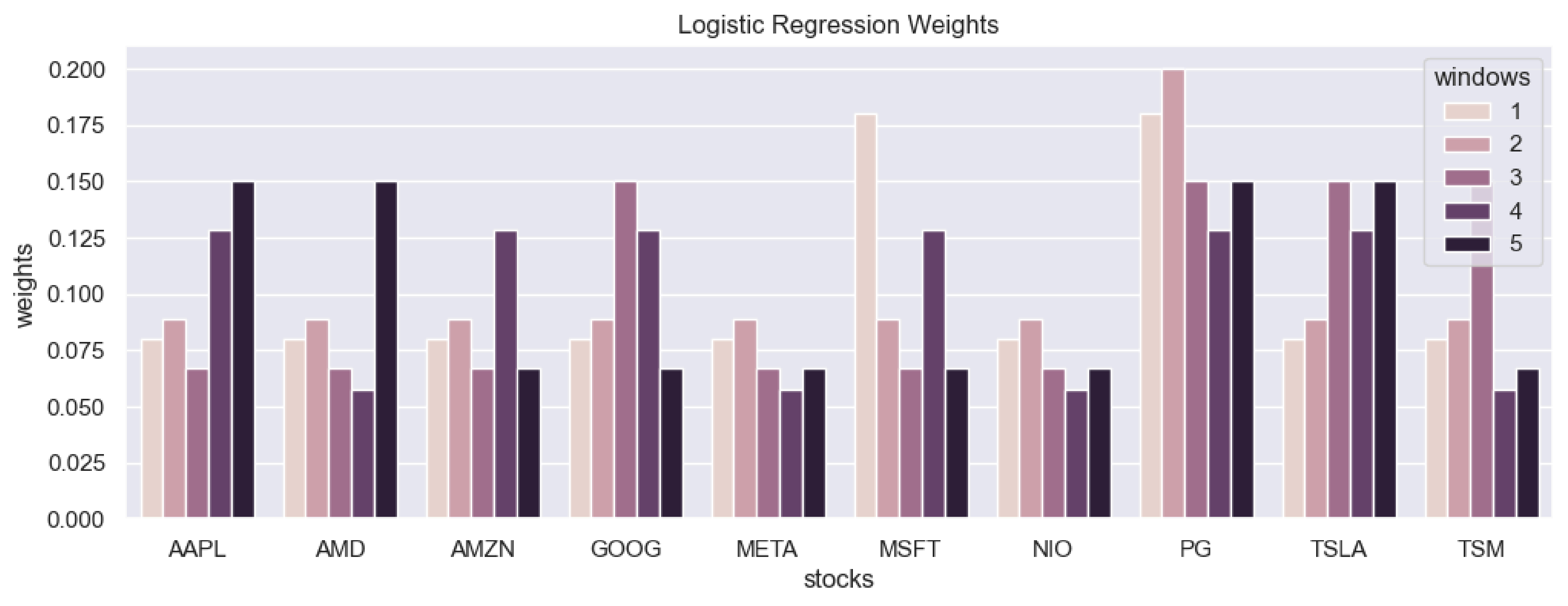
	LR	RF	AB	LR_PCA	RF_PCA	AB_PCA
Annual Return	-0.5377	-0.6298	-0.6405	-0.5564	-0.5437	-0.5739
Annual Volatility	0.3342	0.3221	0.3162	0.3217	0.325	0.3379
Sharpe Ratio	-2.139	-2.9182	-3.0715	-2.3629	-2.2486	-2.3524
Cumulative Return	0.6738	0.571	0.5568	0.6532	0.6665	0.6373
Max Drawdown	-0.3643	-0.3924	-0.3996	-0.3458	-0.3345	-0.3601

ADJUSTING PORTFOLIO WEIGHTS BASED ON PREDICTIONS

Match_weight = 1.5

Weighting changes afterapplying Random Forest





Cumulative Return

Performance Metric

	LR	RF	AB	LR_PCA	RF_PCA	AB_PCA
Annual Return	-0.6077	-0.6353	-0.6199	-0.5917	-0.626	-0.6151
Annual Volatility	0.41	0.427	0.4131	0.4216	0.4597	0.4582
Sharpe Ratio	-2.0733	-2.145	-2.1312	-1.9113	-1.9067	-1.8515
Cumulative Return	0.6121	0.582	0.5982	0.6323	0.6	0.6128
Max Drawdown	-0.3908	-0.4093	-0.3888	-0.3871	-0.4166	-0.4066

SUMMARY

Insights

Performance: some models can slightly improve the annual return and Sharpe Ratio while remaining similar variance.

Prediction Accuracy: LR and RF doesn't have the highest prediction accuracy but they have better performance. The performance has many important factors besides the accuracy.

Limitations and Improvement

Retrieving Sentiment Data from Twitter

- this dataset from Kaggle has limited number of tweets, which may not be enough to find the market sentiment
- use the API from Twitter and select the accounts we want to crawl to retrieve the data.

Formula for Rewarding Matching Sign

- more robust formula to favor the matching case ie. using the exponential of the original weights
- combine the price prediction to adjust the weight or even use the prediction to find the optimal MVP.

Machine Learning Algorithm for Classification

- current fundamental models used to do the classification have limitations on the prediction accruacy
- using some Neural Network Model like the Generative Adversarial Network (GAN)

Shorter Rebalance Period

- the sentiment of the markets may affect the price in a short period of time.
- it is better to implement it as day-trade or 3 day-trade.