1. Strategy Overview

The strategy combines technical indicators with **machine learning models**—specifically **Principal Component Analysis (PCA)** and **Linear Regression**—to predict price movements and make trading decisions. The strategy operates on the SPY ETF (which tracks the S&P 500 index) and aims to generate profits by dynamically entering and exiting positions based on traditional technical signals and predictions from the machine learning models.

This strategy uses the following key elements:

- **Technical Indicators**: RSI (Relative Strength Index).
- **PCA Model**: A machine learning technique that reduces multiple RSI indicators into key components that summarize the most important trends in the data.
- **Linear Regression**: A machine learning model that predicts future returns based on the PCA-transformed RSI data.
- **Feature Extraction**: The key features used for the regression model are various RSI values across different periods.

2. Technical Indicators Used

- **RSI (Relative Strength Index)**:
  - A momentum indicator used to measure the speed and change of price movements.
  - In this strategy, RSI values from periods ranging between 2 and 24 days are used to capture different aspects of market momentum.
  - These RSI values are used as inputs for the PCA model, which then transforms them into key components for predicting future returns.
- **Lookback Period**:
  - The model looks at a 3-day window to calculate future price returns.

The combination of these RSI values provides a solid framework for understanding price momentum and entering positions based on predictions made from the machine learning models.

3. PCA and Linear Regression Model

The **PCA and Linear Regression** models form the core of the strategy's predictive power, both of which are common machine learning techniques. Here's a breakdown of how they work:

PCA and Regression Overview:

- **PCA (Principal Component Analysis)**: PCA is a **machine learning model** used for dimensionality reduction. It transforms a set of correlated RSI values into a smaller set of uncorrelated principal components. This simplifies the input features while retaining the most important information.
- **Linear Regression**: Linear Regression is another **machine learning model** used to predict future returns based on the transformed RSI data (principal components). It models the relationship between the principal components and future returns in a linear way.

- **Why PCA and Linear Regression**: The various RSI values capture different aspects of price momentum. PCA simplifies these inputs into principal components, and Linear Regression is used to predict future returns by learning the relationship between these components and the stock price movements. Both models together form the machine learning backbone of this strategy.

4. Training and Weight Optimization

The goal of training these machine learning models is to learn the optimal relationship between the principal components (from PCA) and future returns (for Linear Regression). The process can be divided into two key phases: training and testing.

Training Phase:

- During training, the PCA and Linear Regression models process historical price data (from 2000 to 2010 in this case) and learn the relationship between the transformed RSI values and future returns.
- The PCA model transforms the RSI data into principal components, and the Linear Regression model is trained to predict future returns based on these components.
- The training process involves the following steps:
    1. **RSI Calculation**: The RSI values for periods 2 to 24 days are calculated based on historical data.
    2. **PCA Transformation**: These RSI values are transformed into 6 principal components.
    3. **Linear Regression Training**: The Linear Regression model learns how the principal components relate to future returns.

Testing Phase:

- Once the machine learning models are trained (i.e., the optimal relationship has been learned), they can be used to make predictions on new, unseen data (from 2010 to 2024 in this case).
- During this phase, the models no longer update their weights. Instead, the Linear Regression model uses the trained coefficients to predict future price movements based on the latest RSI values.

Key Parameters:

- **Number of Components**: 6 principal components are used to summarize the various RSI values.
- **Lookback Period**: The model predicts returns over a 3-day period.
- **Regression Coefficients**: The regression coefficients are learned during training and represent the importance of each principal component in predicting returns.

5. Feature Extraction for the PCA and Regression Model

The machine learning models use multiple features from historical data as inputs:

1. **RSI Values**: The RSI values are calculated for periods between 2 and 24 days.
2. **PCA Transformation**: The RSI values are transformed into principal components that capture the main trends in the data.

3. **Future Returns**: The percentage price change over the next 3 days is used as the target variable for training the regression model.

These features are computed and transformed using PCA, and the resulting components are fed into the Linear Regression model to generate predictions about future price movements.

6. Trading Logic and Rules

Once the PCA and Linear Regression models generate a prediction, the model combines the prediction with signals from the RSI to make trading decisions:

- **Long Entry Conditions**:
  - The Linear Regression model predicts a positive future return (indicating a price increase).
  - If the prediction is positive, the strategy enters a long position in SPY.
- **Long Exit Conditions**:
  - The strategy exits the long position if the predicted return turns negative.
- **Short Entry Conditions**:
  - Although commented out in the strategy code, if implemented, the model could enter short positions based on negative predictions from the regression model.

7. Predictions and Decision-Making

The Linear Regression model's prediction plays a key role in enhancing the decision-making process:

- **Positive Prediction**: Indicates an expected price increase, triggering a long position.
- **Negative Prediction**: Although not currently in use, a negative prediction could be used to trigger a short position.