**Multi-factor Trading Strategy**

1. **Strategy Mechanics**

In the multi-factor trading strategy, we first select 5 least correlated market probability density indicators as factors for each stock. We then employ a rolling regression approach, using these factors to predict stock returns. The predicted returns guide our investment decisions, with weights rescaled to ensure full investment across all tickers. Additionally, we explore machine learning methods like Random Forest and Gradient Boosting Machines to enhance predictive accuracy. This strategy aims to capture diverse market dynamics and optimize returns by leveraging the unique information provided by each selected factor.

1. **Select Factors from MPD Data**

The selection of factors is crucial for capturing different dimensions of market behavior. One approach to factor selection is to choose indicators that are least correlated with each other, thereby ensuring that each factor provides unique information about the market.

The mathematical approach involves calculating the correlation matrix for a set of market probability density (MPD) indicators. The correlation matrix provides a measure of the linear relationship between each pair of indicators. The goal is to select a subset of indicators that minimizes the average pairwise correlation, thereby reducing multicollinearity and improving the model's explanatory power.

For each equity, we select 5 factors for further analysis.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Factor Chosen | | | | | |
| **bac** | mu | skew | kurt | p50 | p90 |
| **citi** | mu | skew | kurt | p50 | prInc |
| **iyr** | mu | sd | skew | kurt | prInc |
| **sp6m** | mu | sd | skew | kurt | p50 |

1. **Construct Trading Strategy**

We employ a rolling regression approach to forecast stock returns based on selected factors. The regression model is fitted on a rolling window of historical data, capturing the relationship between stock returns and factor values over time.

Let represent the matrix of factor values, with each row corresponding to a day in the rolling window , and each column representing one of the five selected factors. Let be the vector of stock returns for the same period. The linear regression model is formulated as:

For prediction, the factor values for the day, are used to forecast the stock return for that day:

In addition to linear regression, machine learning methods such as Random Forest (RF) and Gradient Boosting Machines (GBM) are also explored for their potential to capture nonlinear relationships and improve predictive accuracy.

After obtaining the predicted returns for all tickers, the investment weights for each ticker on day is generated by rescaling the predicted returns such that their absolute values sum up to 1:

1. **Result Analysis**

The table and the image below shows the result of our strategy.

图表

描述已自动生成

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **ticker** | **Return** | **Volatility** | **Sharpe Ratio** | **Skewness** | **Excess Kurtosis** | **VaR (0.05)** | **CVaR (0.05)** |
| **OLS** | 0.05 | 0.22 | 0.22 | -0.47 | 16.05 | -0.02 | -0.03 |
| **RF** | -0.07 | 0.24 | -0.27 | -0.19 | 14.55 | -0.02 | -0.04 |
| **GBM** | -0.07 | 0.24 | -0.31 | -0.16 | 14.78 | -0.02 | -0.04 |

The OLS regression demonstrates a positive annualized return and a Sharpe ratio above zero, suggesting a favorable risk-adjusted return compared to the machine learning methods - Random Forest (RF) and Gradient Boosting Machines (GBM). Both RF and GBM exhibit negative annualized returns and negative Sharpe ratios, indicating underperformance on a risk-adjusted basis.

The graphical representation of the strategy's profit and loss (PnL) over time further illustrates the disparity in performance. The OLS method achieves a progressively increasing PnL trajectory, while the machine learning methods do not exhibit the same degree of success, with their PnL lines remaining below that of OLS and displaying higher volatility.

One critical factor to consider is the frequency mismatch between the MPD data and the stock return data. The MPD data is monthly, providing a low-frequency view of market conditions, while the stock return data is daily, necessitating a high-frequency predictive model. Machine learning methods, particularly RF and GBM, are generally more complex and may overfit the noise within high-frequency data while failing to capture the underlying patterns adequately when trained on lower-frequency features. This can result in models that do not generalize well out-of-sample, leading to poor real-world performance as observed.

In conclusion, for the given strategy, where MPD data is less granular than the return data, simpler models like OLS outperform their machine learning counterparts. This suggests that machine learning methods may not be suitable for this strategy, as they require high-frequency data to learn and adapt effectively. As a result, we will introduce pair trading strategy in the following part.