Haibo Xie
Data Engineering

Final Projects

Introduction and Purpose

In the vast financial market, the number of investment options available easily overwhelm retail investors. Often, their choices are based on intuition, random guess, or fragmented and outdated information. Consequently, the portfolios they construct are usually far below the efficiency frontier, characterized by higher volatility and lower expected returns, which contradicts their typically high risk aversion feature. While thorough due diligence is ideal for value investors focused on long-term investment horizons in specific assets, it requires substantial financial expertise and is time-consuming, often rendering it inefficient from a productivity standpoint. To address the emotional and unsophisticated investment decisions, data-driven analysis in the realm of machine learning is the best alternative choice to gain insights through effective communication with data. The purpose of this project is to classify assets into appropriate classes and construct efficient trading strategies.

Techniques implement and procedure analysis

In terms of efficiency, classification as one of the supervised machine learning techniques can provide a comprehensive means of time-sensitive categorization of assets based on their selected features. The labels being assigned to assets are in three classes: forecastable, undefined, and unforecastable.

So how do we really classify?
The first step is to construct a weighted average base portfolio which consists of several most liquidable assets with distinct features as representation of a simple market portfolio. Run a regression test on the aggregate features of the base portfolio and identify mean squared errors of asset prices, then assign labels based on the scale of this measurement. The larger of the MSE means that the less accuracy of the base portfolio regression, in other words, it should be assigned as non-forcastable, and vice versa for the smaller the MSE, the more power of forecastability of this regression. Assets with in-between values of MSE will be classified as undefined.

The second step, training the classification on features of the base portfolio and predicting labels on objects in the new dataset. The result of classification scales down our investment pool, which is then handed to the decision maker of choosing suitable candidates from the pool to construct different types of trading strategies.

Now moving on to the trading strategy implementation

The strategy we select is the classical long-short strategy, which is to the point of fully capturing the opportunities of maximizing profit by both taking long positions on the underpriced assets and short positions on overvalued assets, meanwhile diversifying the risk, which align with the goal of efficient portfolio, which is to minimize the standard deviation with a higher expected return.

The assets of long position and short position should ideally have a high negative correlation with each other. As we know that no perfect negative-correlation assets exist in the financial world, a moderate to a relatively high level correlation is sufficient. In the FX market, the USD/JPY and GBP/USD have a historical negative correlation (5-year daily price correlation: -0.56). To identify which to long and which to short, for simplicity, we run a simple linear regression model on each trading currency pair, and the regressor is the timing factor representing embedded momentum characteristics. Currency pairs with the higher coefficient of the timing factor will be expected to have a larger magnitude of momentum within a specific period.

Lastly we open up the positions of long and short with the L-S ratio equal to 1 as mimic of the initial investment capital, and keep making reinvestment without rebalancing the portfolio on a defined frequency within the investment horizon. At the end of this investment, the result of output will be used to evaluate the overall performance of the project.

Intermediate Results

Portfolio regression (best model: linear regression)
MSE of base currency pairs to the portfolio regression
MSE for EUR/USD: 0.0002079 Undefined
MSE for GBP/CHF: nearly a 0 Forecastable
MSE for USD/CAD: 0.0004121 Non-forcastable

Classification (best model; Naive Bayers)

EUR/CHF:prediction_label
FORECASTABLE 87.096774
UNDEFINED 12.903226

EUR/CAD:prediction_label
NON-FORECASTABLE 100.0

GBP/EUR:prediction_label
FORECASTABLE 100.0

GBP/USD:prediction_label
NON-FORECASTABLE 100.0

GBP/CAD:prediction_label
NON-FORECASTABLE 100.0

USD/CHF: prediction_label
FORECASTABLE 100.0

USD/JPY:prediction_label
UNDEFINED 100.0

From the classification results of the target currency pools, we can identify labels of trading currency pairs. USD/JPY result in 100% undefined and GBP/USD result in 100% non-forcatable, which deviates from the historical intuition of the tradability of two currency pairs. Several potential contributions to this scenario:

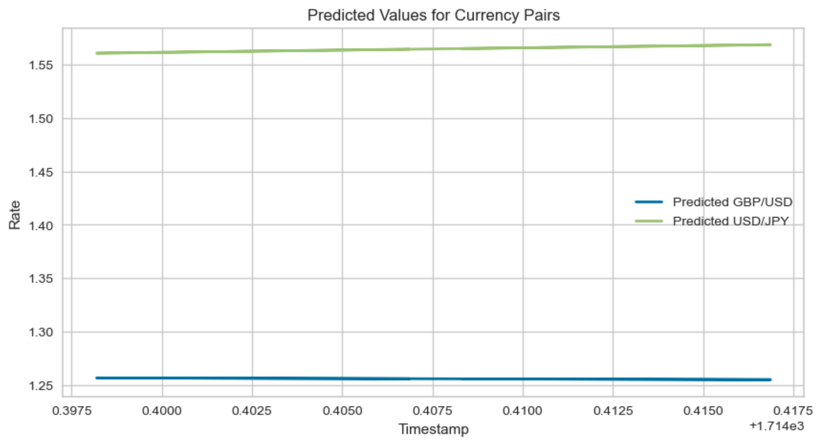
> One is that relative ranking to determine the forecastability is not valid, since all base currency pairs have a low MSE, which lacks the diversity of constructing this base portfolio and entails the later on problematic classification.
>
> Another factor is that the regression is performance on a small-scale dataset, which needs to expand the size of the dataset to truly reflect a more powerful portfolio regression.
>
> Another scenario could happen if the currency pair really changes its behavior and needs further analysis.

Regression analysis for long and short position
GBP/USD

Model	MAE	MSE	RMSE	R2	RMSLE	MAPE
Linear Regression	0.0002	0.0000	0.0002	0.8189	0.0001	0.0001

USD/JPY

Model	MAE	MSE	RMSE	R2	RMSLE	MAPE
Linear Regression	0.0009	0.0000	0.0011	0.8109	0.0004	0.0006



Coefficient of GBP/USD is higher than USD/JPY within the previous 5 hours. To better visualize the regression, from the graph, there is an upward trend of GBP/USD, and a slightly downward trend of USD/JPY. This infers that within the past 5 hours, GBP/USD is trending upward, USD/JPY is in a decreasing movement. Thus, on the assumption of the continuity of this momentum, I decided to long GBP/USD and short USD/JPY.

Final Trading result

	USD/JPY_rate	GBP/USD_rate	L_position	S_position	LS_ratio	LS_ratio_1	LS_ratio_2
2024-04-29 22:56:02.895286	156.7420	1.25485	1.25485	156.742	1.000000	NaN	NaN
2024-04-29 22:56:03.591423	156.7420	1.25485	1.25485	156.742	1.000000	NaN	NaN
2024-04-29 22:56:04.202788	156.7480	1.25485	1.25485	156.742	1.000038	NaN	NaN
2024-04-29 22:56:04.817146	156.7460	1.25484	1.25485	156.742	1.000033	NaN	NaN
2024-04-29 22:56:05.457443	156.7485	1.25484	1.25485	156.742	1.000049	NaN	NaN
...	...	...	...	...	...	...	...
2024-04-30 01:55:58.957244	156.7970	1.25339	1.25485	156.742	1.001516	1.001516	1.001516
2024-04-30 01:55:59.637441	156.7970	1.25339	1.25485	156.742	1.001516	1.001516	1.001516
2024-04-30 01:56:00.398386	156.7965	1.25340	1.25485	156.742	1.001505	1.001505	1.001505
2024-04-30 01:56:01.419671	156.7970	1.25340	1.25485	156.742	1.001508	1.001508	1.001508
2024-04-30 01:56:02.214498	156.7975	1.25340	1.25485	156.742	1.001511	1.001511	1.001511

Over a 3 hour investment horizon

First investment:
Start: 1
End: 1.001511
Return: 0.1511%

Reinvestment 1
Start: 1.000435

End: 1.001511
Return: 0.1076%

Reinvestment 2
Start: 1.000459
End: 1.001511
Return: 0.1052%

Total return: 0.3639%

Arithmetic average return: 0.1213%

In reality, due to capital limitations, reinvesting at such a frequency is impractical. Therefore, I will use a return of 0.1511% over a 3-hour period as the performance metric. Converting this to a daily return yields 1.22%, and annualizing it results in a significantly high return. However, this assumes an average return of 0.1511% every 3 hours without accounting for service and transaction fees. Overall, these results are a promising indication that the methodology implemented in this project is both feasible and effective.

Improvements:

1. Base portfolio correlation cannot simply average out the constituents correlation to the exogenous variables, which is not accurate enough, instead, I should use the weighted average portfolio price to compute correlations
2. Base portfolio should construct by more currency pairs to be more diversified
3. Rebalanced trading portfolio over time to prevents the distortion of strategy
4. The enter point and existing point should be determined by the measurement of FD and volatility, and other factors that can estimate the direction of movements