1 Summary

This project aims to propose a signal for a broad-based ETF. The ETF we choose is PDR S&P 500 ETF trust (SPY). Leveraging the spirit of strong learner in machine learning, our general idea is to ensemble predictions for individual stocks to produce a signal for SPY prediction. Furthermore, we specifically put forward a ETF flow signal that could be used in individual stock return prediction.

2 ETF chosen

SPY is an ETF designed to track the S&P 500 stock market index. We choose it because it has large AUM. Based on data up to Dec 20 2023, SPY has an AUM of $487,060.36 M. Our ensemble signal will work better on a ETF covering a broad range of stocks with large AUM.

3 Idea 1: individual stock prediction ensemble

3.1 Idea

We first build predictions on a group of individual stocks using traditional alpha factor. Then we ensemble these predictions according to each stock's weight in the index to produce a signal for SPY prediction.

3.2 Rationale

SPY is a single instrument. If we directly build a signal to forecast SPY, we will have only a few data points. Models can discover limited information from a small dataset. Such a signal will be less robust in outsample prediction.

However, if we focus on a basket of stocks which are related to this index, we will have a much larger dataset. Predicting a basket of stocks will be easier than predicting a single SPY. Furthermore, the ensemble prediction for SPY will have even lower bias and variance compared to prediction on single stocks.

3.3 Implementation

3.3.1 Stock Selection

We will first include all components in S&P500. Beyond that, we will include another 300 - 500 stocks according to the selection criteria of S&P500, including relatively large market cap, relatively high liquidity, positive earning recently. These additional stocks have similar properties as those in the S&P500's basket, and thus would provide relevant information for prediction. While the prediction for these additional stocks will not be used to ensemble the final SPY prediction, more data allows us to train a more robust and accurate model.

3.3.2 Y label Selection

We forecast return single stocks, instead of their price, since price is not stationary.
In terms of prediction horizon, there is a trade-off. Prediction on long horizon will be more difficult due to information decay in time. But short horizon prediction and more frequent trading will lead to higher turnover and commission fee. Currently, we assume daily prediction. We will need data analysis to measure the actual cost and benefit and to decide the prediction horizon.

3.3.3 Signals

We will use basic signal for stock price prediction, including signals built on price volume data, fundamental data, alternative data. We don't have high requirement on each single signal, since the spirit of our idea is ensemble.
Specifically, we put forward a **ETF flow signal**, which will be introduced in next section.

3.3.4 Model Selection

The data size of our model is roughly 10 years $*$ 252 trading days $*$ 1000 stocks = 2.5 M rows of data. We will choose LightGBM to ensemble the signals, for its superior performance in prediction accuracy and efficiency on dataset with 1M+ rows.

3.3.5 Model Training

First, rolling windows will be applied to deal with the problem of market regime change. Second, in each window, we apply purged K-fold cross validation to most efficiently utilize data. We will train

a model on each (K-1) folds and a final model on the whole training set, and ensemble the (K+1) models to produce the final prediction. Different from regular K-fold strategy, purged K-fold strategy removes several days' data between every two folds to avoid data leakage through rolling features.

3.3.6 Prediction Ensemble

We calculate a weighted average of predictions on single stocks based on their weight in the index. This is the prediction for SPY return.

4 Idea 2: ETF flow

4.1 Idea

This is a signal for prediction individual stock return, which can be used in building the signal in Section 3.

$$signal_{i,t} = ewma(normalize(\sum_{\text{all ETFs}} w_{i,j,t} * flow_{j,t})) \tag{1}$$

for day t, stock i, ETF j

4.2 Rationale

We define ETF flow as the money move in and out of a ETF. Every day, ETF creations, ETF redemptions, and index rebalance, will all lead to trading on underlying stocks. Since ETF investment is passive, we believe there will be a reversal effect on the price of these underlying stocks.

4.3 Data Support

From Figure 1 provided by BlackRock, approximately 5.2% of trading volume in U.S. equities has been attributable to ETF activity, higher than the number in other markets. Furthermore, once there are index rebalance, data suggests that index funds could account for over 40% of all trading on that day. These statistics suggest that ETF flow is probably large enough to cause the subsequent reversal in individual stock prices.

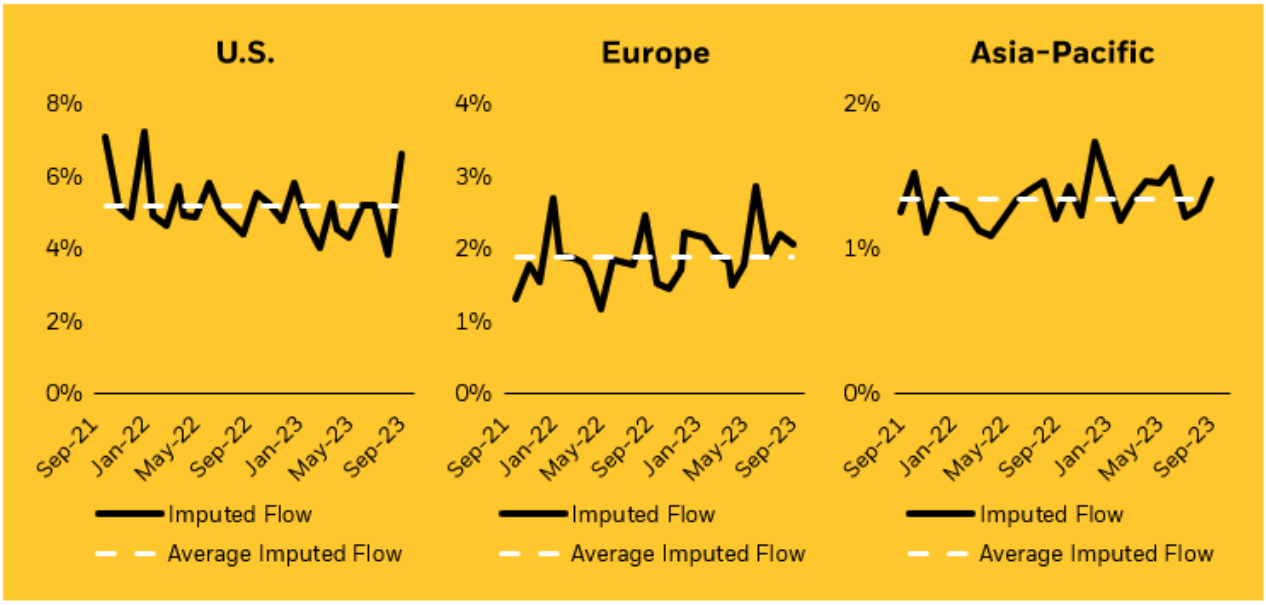

Chart description: Line charts showing both the total and average imputed flow in the U.S., Europe, and Asia-Pacific. Imputed flow is an estimation of how stock trading is generated by ETF inflows and outflows. The charts show that imputed flow is below 5.3%, on average, in all regions.

Figure 1: Percentage of stock trading as a result of ETF flows (Source: BlackRock Market Insights)

4.4 Implementation

- collect the daily net flows data of all major ETF in the market (data source: ETF net flows data link)
- calculate the corresponding net flow on each individual stocks based on the holdings and net flow of ETFs
- normalize the signal and apply exponentially moving average on the signal to reduce noise.