Finance Project Report

Fei Du (fd2547), Shiyun Chen (sc5324), Chanjin Park (cp3366)

0 Introduction

Invesco QQQ Trust Series 1[1] includes 100 of the largest nonfinancial companies mostly in the US, which is based on the Nasdaq-100 Index. Top ETF holdings include Apple, Microsoft, Nvidia, and etc. This report will provide a long-only trading strategy where the investment horizon is one day from comprehensive data analysis and modeling techniques.

1 Data Collection and Preprocessing

1.1 Raw Data

The raw QQQ data is scraped from Yahoo Finance[2], including open price, close price, high, low from '2023-01-01' to '2024-12-06.' The dataset also contains Nasdaq-100 EPS for better prediction, where the data source is Gurufocus[3].

1.2 Other Indicators

For better performance of the model, the dataset includes various indicators implying the trend of the price and market context:

- **Daily return**: The daily return captures the momentum of the price. This is useful to determine the short-term trend of this ETF.
- **Weekly volatility (5 days rolling data)**: The weekly volatility measures the price variability and indicates the risk of stock.
- **Weekly SMA & EMA (5 days rolling data) and Monthly SMA & EMA (21 days rolling data)**: Both SMA and EMA present the trend of the stock price. By including both SMA and EMA, the model can consider the overall trend and the trend prioritizing current value. By including both weekly and monthly MA, the model can examine the performance of price from a short-term view and a broader perspective.
- **Nasdaq-100 EPS**: QQQ is based on Nasdaq-100, so understanding the situation of Nasdaq-100 is crucial for predicting the performance of QQQ. Nasdaq-100 EPS represents the earning growth, which might be helpful for deciding the profitability of QQQ.

[1] https://www.invesco.com/us-rest/contentdetail?contentId=3a48e01e98630410VgnVCM10000046f1bf0aRCRD
[2] https://finance.yahoo.com
[3] https://www.gurufocus.com

- **Nasdaq-100 P/E**: Nasdaq-100 P/E, which is calculated based on the Nasdaq-100 price and EPS, represents the valuation level of this stock, and thus influences the trading decision of QQ by understanding whether the stock is overvalued or undervalued.
- **S&P 500**: S&P 500 includes more companies and reflects a broader context of the market.
- **S&P 500 daily return**: S&P 500 daily return can be a valuable indicator of the broader market trend in the short-term.
- **VIX (Volatility of the market)**: VIX measures the investors' sentiments to the market and can be a reference when making our own decisions.
- **Interest Rate (10-year treasury yield)**: The interest rate indicates the monetary policy that will influence the equity market.
- **Oil (Crude oil price)**: The oil price also indicates the economic concerns globally, which is highly associated with the equity market.

1.3 Data Manipulation

The dataset is created by shifting all the representative indicators except 'Open' up for 1 day. This makes sure that the model will not foresee information about the stock improperly. Before delivering the data to the model, there is also a standardization step used to normalize the data for each feature. The purpose is to ensure all features contribute equally to the model.

2 Methodology

2.1 Trading Strategies

a. Buy an ETF at open*(1 - 0.1%). We don't trade if this cannot be reached;
b. Sell ETF when the price reach the stop profit limit which is (open + 0.5*previous ATR). Otherwise, we sell at the close price of the day.

2.2 Label Defined Methods

Based on the strategies listed above, the true labels are defined by:

a. If the price during the day never drops to Open * (1 - 0.1%), the true label is 0, meaning we should not trade;

b. If the price during the day drops to Open * (1 - 0.1%) and later reaches the stop-profit limit (with the stop-profit occurring after the market entry), the true label is 1, meaning we should trade because there is a profit opportunity;

c. If the price during the day drops to Open * (1 - 0.1%) but never reaches the stop-profit limit, comparing the closing price and the entry price. If Close - Entry Price > 0, it indicates that we profit by the end of the day, so the true label is 1 (we should trade). Otherwise, the true label is 0 (we should not trade).

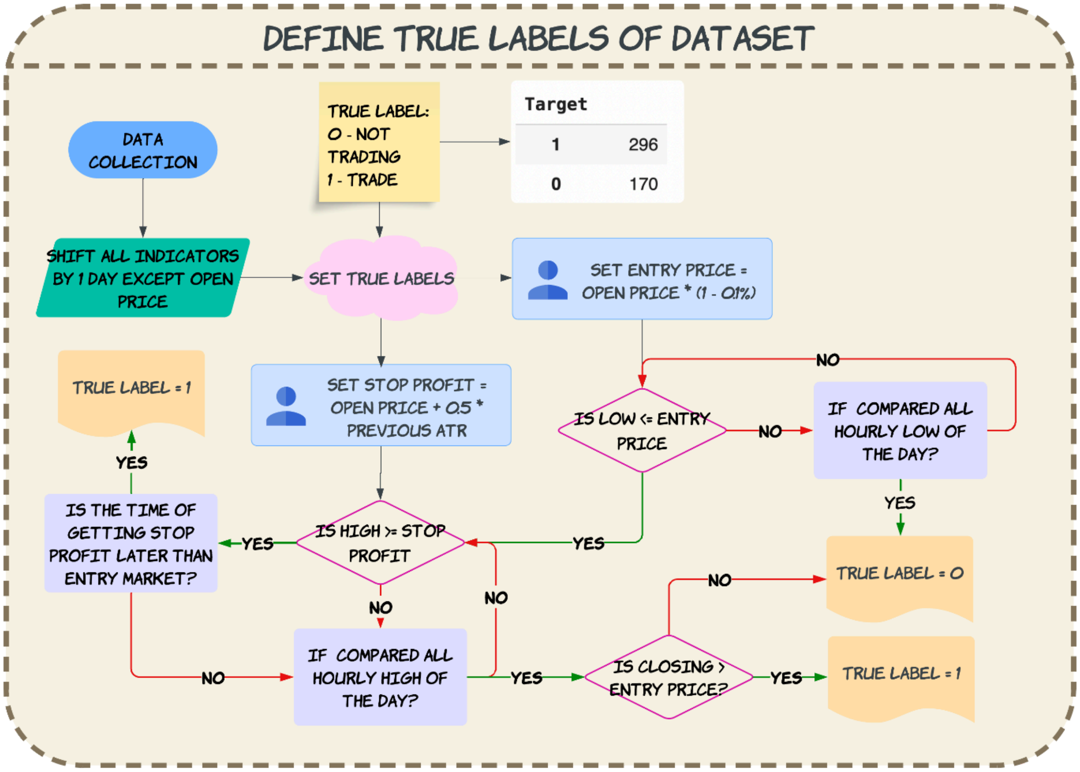
DEFINE TRUE LABELS OF DATASET
DATA COLLECTION
TRUE LABEL: 0 - NOT TRADING 1 - TRADE
Target
1 296
0 170
SHIFT ALL INDICATORS BY 1 DAY EXCEPT OPEN PRICE
SET TRUE LABELS
SET ENTRY PRICE = OPEN PRICE * (1 - 0.1%)
TRUE LABEL = 1
SET STOP PROFIT = OPEN PRICE + 0.5 * PREVIOUS ATR
IS LOW <= ENTRY PRICE
NO
IF COMPARED ALL HOURLY LOW OF THE DAY?
NO
YES
YES
IS THE TIME OF GETTING STOP PROFIT LATER THAN ENTRY MARKET?
YES
IS HIGH >= STOP PROFIT
YES
TRUE LABEL = 0
NO
NO
NO
NO
IF COMPARED ALL HOURLY HIGH OF THE DAY?
YES
IS CLOSING > ENTRY PRICE?
YES
TRUE LABEL = 1

Fig 1. Flowchart of Trading Strategies

3 Modeling

We have tried multiple models for this part including time-series models like Prophet and deep learning models like LSTM. It turns out that XGBoost has the best performance. The reason may be that the amount of data is not enough for deep learning models to learn the pattern, and for this kind of amount of data, XGBoost is a better option because it can catch the complex nonlinear pattern exists in the data and allow people to adjust multiple hyperparameters to improve the model performance.

Table 1. Comparison of LSTM and XGB Model

Methods	LSTM	XGB
Accuracy	0.628571	0.642857
Usage	Predict stock price for future days	Classify whether to trade or not
Dataset Requirement	Need large amount of historical data	Efficient with small datasets

There are 466 data points in total, and the model takes the first 70% data as training set and the last 30% data as testing set. After tuning hyperparameters, the model reaches an in-sample accuracy of 97.65% and an out-sample accuracy of 64.29%. The best hyperparameters are colsample bytree = 0.8, learning rate = 0.01, max_depth = 10, n estimators = 50, and subsample = 0.8. The features are open, previous daily return, previous volatility, previous weekly SMA&EMA, previous monthly SMA&EMA, previous Nasdaq-100 EPS, previous Nasdaq-100 P/E, previous S&P 500, previous S&P 500 daily return, previous VIX, previous yield, previous oil price, previous close, previous high, previous low, and stop profit limit.

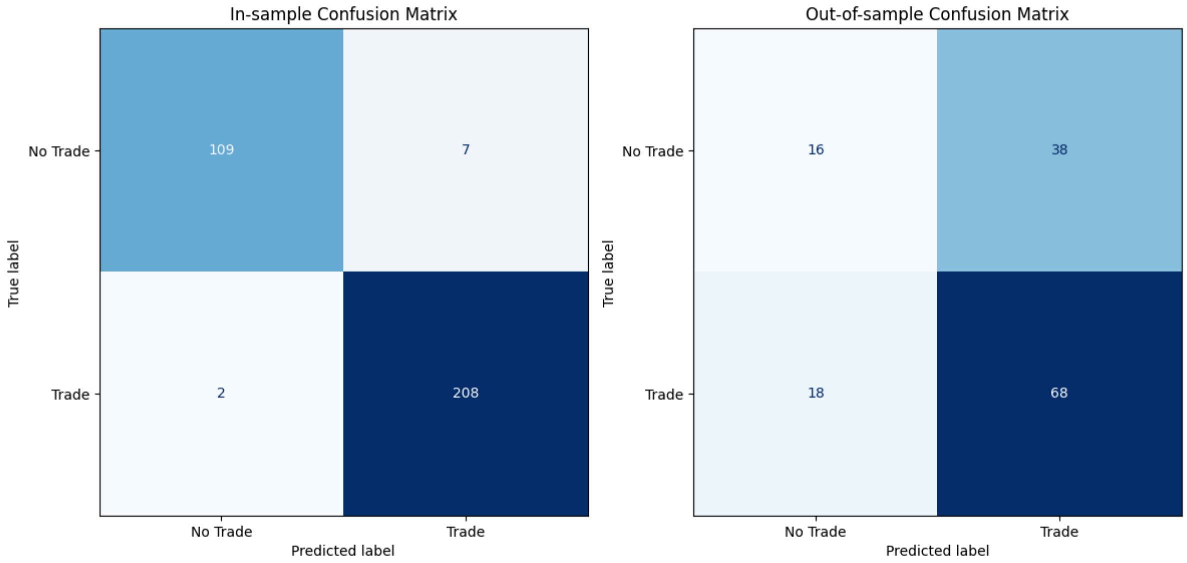

Fig 2. Performance of In-Sample and Out-of-Sample Data

The confusion matrix indicates that the model takes an aggressive approach to trading decisions. It demonstrates higher recall but lower precision for cases labeled as ‘trade’, emphasizing the identification of as many trading opportunities as possible, even at the cost of false positives. Conversely, it exhibits higher precision and lower recall for cases labeled as ‘not trade’, showing greater caution in predicting no-trade scenarios, which may result in missed opportunities. Financial market data is often influenced by random noise and spurious correlations, so it poses a risk of overfitting, where the model captures patterns in the training data that fail to generalize to new data. Despite these challenges the model achieves an accuracy of 64.29%, providing a statistical advantage over random guessing (50%). When implemented consistently across numerous trades with sound risk management, this edge can lead to profitability. This model is thus appropriate for users who care about opportunity loss and want to catch chances to trade and make profit as much as possible, even if trading may lead to loss.

4 Conclusion and Future Steps

As a conclusion, the model works well on predicting whether the customer should go long for QQQ given the open price of that day and other related information from days before. If customers enter the market, they will make profit in around 65% of the cases.

The future steps of this market include:

- Exploring longer investment time spans, such as 2-day or 3-day horizons, to determine which duration yields better prediction accuracy and trading performance;
- Calculating the P&L under the current strategy and model (include opportunity cost);
- Offering portfolio suggestions given the current P&L and model accuracies or likelihood of each decision;
- Analyzing the seasonal patterns of this ETF using time series models and using the patterns to modify our trading decisions.