Alpha Generation by Machine Learning

Muya Liu
Xuan Liu
Xinyue Cao

Contents

1.Overview 1

2.Data Preparation 1

2.1 return 1

2.1.1 Basic information 1

2.1.2 Data collection： 1

2.1.3 Data preprocessing 1

2.2 factors 1

2.2.1 concept of factors 1

2.2.2 data acquisition 3

3.Alpha generation 4

4. Machine Learning model 5

4.1 model selection 5

4.2 parameter adjustment 5

4.3 model fusion 6

5. result 6

5.1 basic factor analysis 6

5.1.1 Factors correlation 6

5.1.2 Rank correlation 7

5.1.3 SHAP value 8

5.2 The relationship between alpha and factors in 4 models 9

5.2.1 coefficients of linear models 9

5.2.2 feature importance of boosting models 9

1.Overview

This project builds an **alpha generation model** for Nasdaq stocks using machine learning. Our goal is to find variables explaining returns unrelated to market volatility. Utilizing historical data from **2019 to 2023**, we first calculate alpha using the expected returns derived from the **Fama-French Five-Factor Model**. Next, we use different **machine learning models** such as to fit the data and optimize their parameters. Then we use **Soft Voting** to combine these models. Finally, we identify key factors influencing alpha by evaluating their explanatory power.

2.Data Preparation

2.1 return

2.1.1 Basic information

Targeted Market: NASDAQ market

Time Range: 2019-01-01 to 2023-12-29

Data Source: CRSP / Annual Update / Stock / Security Files / Daily Stock File

2.1.2 Data collection：

Step1: Download all NASDAQ sector data.

Step2: Eliminate stocks with abnormal yields (non-numeric values) and stocks without yield data on the last trading day in the time range. These are likely to be stocks that have been delisted before 2024 which have no significance for alpha generation research.

Step3: Calculate the total trading volume of each stock and filter out the top 500 stocks in terms of trading volume.

Step4: Convert the data format, set column to ticker, index to date, and value to return dataframe, so that we can process the data of each stock separately.

2.1.3 Data preprocessing

Step1: We firstcheck the missing data. According to the data statistics, the missing values of each stock are all in the first half of each column of data. That is to say, the reason for these missing values is that the stock was not listed on this date. Therefore, we should not fill the forward data.

Step2: In order to avoid the influence of outliers on the model, we use winsorize processing to shrink the tail by 1%.

2.2 factors

2.2.1 concept of factors

Factor models are key tools for understanding and predicting asset returns. We categorize the factors into 7 groups, which are market factors, fundamental factors, moving average factors, overlap factors, momentum factors, volume and volatility.

The following table is an overview of the factors we use in the model.

Factor Type	Amount	Main Factors
Market	5	Mkt-RF, SMB, HML, RMW, CMA
Fundamental	33	AbsAccrual, DIV_P, TaxExp, CURRENT, FCF_P, CEQG, momaccel⋯
Moving Average	11	SMA, WMA, TEMA, MAMA
Overlap	10	BBANDS_upper, BBANDS_up, HTITREND, SAR, SAR_norm⋯⋯
Momentum	24	MACD, STOCH, WILLR, CMO, MFI ⋯⋯
Volume	5	Chaikin A/D Line, On Balance Volume, Volume-Price Trend, Volume Weighted Adjusted Price
Volatility	3	ATR, ATR_normalized, NATR

(1) Market Factors

Market factors are key drivers of asset returns, capturing overall market risk and return dynamics. Here we use the 5 factors in Fama French's model as the market factor.

(2) Fundamental Factors

Fundamental factors are based on a company's financial data, revealing the intrinsic characteristics of assets. Here are a few examples:

Cash Dividends: Profits distributed to shareholders in cash, indicating financial stability and the ability to generate consistent cash flow.

Current Ratio: Current assets divided by current liabilities, used to measure short-term liquidity.

Quick Ratio: A stricter measure of liquidity, considering only the most liquid assets.

Free Cash Flow: Cash remaining after operational expenses, available for investment or shareholder distribution.

Operating Profitability: The ratio of operating income to total assets, reflecting the profitability of core business operations.

(3) Moving Average Factors

Moving average factors smooth price fluctuations and identify trend changes. The key types include:

Simple Moving Average (SMA): Averages prices over a specific period to reduce noise.

Weighted Moving Average (WMA): Assigns higher weights to recent data, making it more responsive to recent changes.

Triple Exponential Moving Average (TEMA): Combines multiple exponential moving averages to reduce lag.

Adaptive Moving Average (MAMA and FAMA): Dynamically adjusts to market cycles.

(4) Overlay Factors

Overlay factors use data from various time windows to evaluate asset volatility and trend strength. Key indicators include:

Bollinger Bands: A technical indicator based on a moving average and standard deviation to measure price volatility.

Hilbert Transform - Instantaneous Trend:

Hilbert Transform detects trends or cycles in price signals:

HTITREND: Reflect the instantaneous trend direction.

HT_norm: Normalized Hilbert indicator for comparing trend strength across assets.

(5) Momentum Factors

Momentum factors capture the continuation of asset price trends, forming the basis of momentum trading strategies. Key factors include:

pDM (Plus Directional Movement): Measures upward price movement compared to the previous period.
mDM (Minus Directional Movement): Measures downward price movement compared to the previous period.
APO (Absolute Price Oscillator): The difference between long-term and short-term moving averages, used to identify trend reversals.
AROON_up / AROON_down: Indicates the proportion of time prices have reached highs or lows in a specific period, reflecting trend strength.

(6) Volume Factors

Volume factors analyze trading activity to evaluate market momentum, trend strength, and fund flows. Key factors include:

AD (Chaikin A/D Line): Measures money flow trends using price and volume data, indicating trend strength.
ADOSC (Chaikin A/D Oscillator): Tracks short- and long-term money flow changes to confirm trends or signal reversals.
OBV (On Balance Volume): Combines volume with price direction to assess whether volume supports price trends.
VPT (Volume-Price Trend): Analyzes the relationship between volume and price changes to track fund flows and trend strength.

(7) Volatility Factors

Volatility factors measure the intensity of asset price fluctuations, reflecting market uncertainty and sentiment. Key factors include:

ATR (Average True Range): The average range of price movement over a specific period, indicating the magnitude of price volatility.
NATR (Normalized ATR): ATR as a percentage of the current price, allowing comparison across different assets or time periods.

2.2.2 data acquisition

(1) market factors

Data Source: Ken's French / Fama/French 5 Factors (2x3) [Daily]

We chose the Fama French five-factor model for alpha generation. We can directly grab the data from Ken's French website.

(2) transaction related factors

Data Source: CRSP / Annual Update / Stock - Version 2 (CIZ) / Daily Stock File

The transaction-related factors are mainly derived from feature engineering based on the close price, high/low price volume of the stock, so we need to download data related to daily transactions on CRSP.

Time Range: 2018-01-01 to 2023-12-29

Since most of the factors need to use the data previous to the time period we need in the calculation, we download some earlier data.

Calculation function:

There are a lot of encapsulated calculation functions in the python library TA-Lib. We generate all factors in the time period for each ticker separately.

Data Integration:

After obtaining all the factors for each ticker, we merge these factors and the return data together. To make the model results more predictive we match the alpha of t+1 with the factor of t in order.

(3) fundamental factors

The data acquisition and processing of fundamental factors are similar to that of trading related factors. We need to match today’s alpha/return with previous factors.

Basic information:

Data Source: Wharton Research Data Service / Factors by WRDS

Time range: 2018-12-30 to 2023-11-30

Data integration:

to the point we focus on in this part is that the fundamental factors in WRDS are monthly data, while the returns are daily data. Same to the trading related factors, we need to match the fundamental factors of the previous period. Here we match the alphas of this month with the fundamental factors in last month.

Missing value processing:

Step1 For data in every column, if its missing values of a single factor accounted for more than 30%, we delete this part of the data directly.

Step2 In the remaining data, select all the data without missing data.

Step3 In order to simplify the procedure, here we select the stocks that have all date data in these 5 years.

3.Alpha generation

The key issue of our project is to figure out where the excess return comes from. Here we assume there are two parts which will have impacts on the excess return-the market and the feature of the stock itself. For the market impact, we use the Fama French five-factor model to calculate alpha, and we construct the machine learning based alpha generation model.

$$\begin{aligned} r_{i,t} - r_{f,t} = \alpha_t + \beta_{1,t}(r_m - r_f) + \beta_{2,t} \times SMB + \beta_{3,t} \times HML \\ + \beta_{4,t} \times RMW + \beta_{5,t} \times CMA + \varepsilon_t \end{aligned} \tag{1}$$

From the mathematical formula of the Fama French 5-factor model shown above, we define alpha as the intercept which cannot be explained by the market. In our model, we use the rolling OLS (rolling window = 30) to calculate the alphas for every ticker.

From a global perspective, the generation of alpha is a two-way process. From the perspective of historical data and model training, we need to obtain this alpha through regression using the Fama French 5 model, and then use it as the dependent variable of the alpha generation model to train the model. But if we want to do the future forecasts and to give the investment suggestions, it is a totally inverse procedure, we get this part of factors first and we use the machine learning to predict the alpha.

4. Machine Learning model

4.1 model selection

For model selection, we use five models, ElasticNet, Lasso, Ridge, Huber Regressor, and RANSAC Regressor from machine learning linear regression, to compare with lightgbm, xgboost, and catboost from boosting algorithm. The results are as follows:

Model	Type	MSE	Running time
ElasticNet	Linear	1.9827	71s
Lasso	Linear	2.0189	18.5s
Ridge	Linear	1.6758	1.6s
Huber Regressor	Linear	2.1682	22.5s
RANSAC Regressor	Linear	12.7320	20.2s
LightGBM	Boosting	0.9487	24.7s
XGBoost	Boosting	1.1118	324.6s
CatBoost	Boosting	0.9628	250.9s

In the linear regression, we found that the Ridge model and the ElasticNet model had smaller MSE and shorter running times. Thus, we continue to adjust parameters for these two models. In the Boosting models, XGBoost is a good fit but the running time is too long, so we only parameterize the other two Boosting models.

4.2 parameter adjustment

In this part, we use grid search to adjust the parameters. Since our alphas are time series data, we use TimeSeriesSlpit instead of traditional KFold or stratifiedKfold for cross validation. Here are the results of parameter adjustment.

Model	Parameter	MSE Before	MSE After
ElasticNet	{'alpha':0.01,'l1_ratio': 0.1}	1.9827	1.7406
Ridge	{'alpha': 10, 'Max_iter': 500}	1.6758	1.6757
LightGBM	{'learning_rate': 0.1, 'max_depth': 3, 'num_leaves': 30}	0.9487	0.9166
CatBoost	{'depth': 7, 'iterations': 1500, 'learning_rate': 0.01}	0.9628	0.8676

The result shows that the accuracy of the model has been improved to a certain degree after parameter adjustment.

4.3 model fusion

As every machine learning algorithm has its own advantages and disadvantages, we use the voting classifier, which is a typical algorithm in ensemble learning, to integrate the fitting results of the above three models. Under the voting classifier, we choose the Soft Voting which can find the probability mean of all the classes. Thus, the class with the highest probability would be adopted as its final prediction result.

The following table shows the results of soft voting for different weights of the model:

model	weights					
xgboost	1	0	0	1	3	1
lightgbm	0	1	0	1	2	2
catboost	0	0	1	1	1	3
MSE	1.1118	0.9487	0.9628	1.114	1.263	1.2292

However, we find that not all the voting weights can improve the accuracy of the model.

5. result

5.1 basic factor analysis

5.1.1 Factors correlation

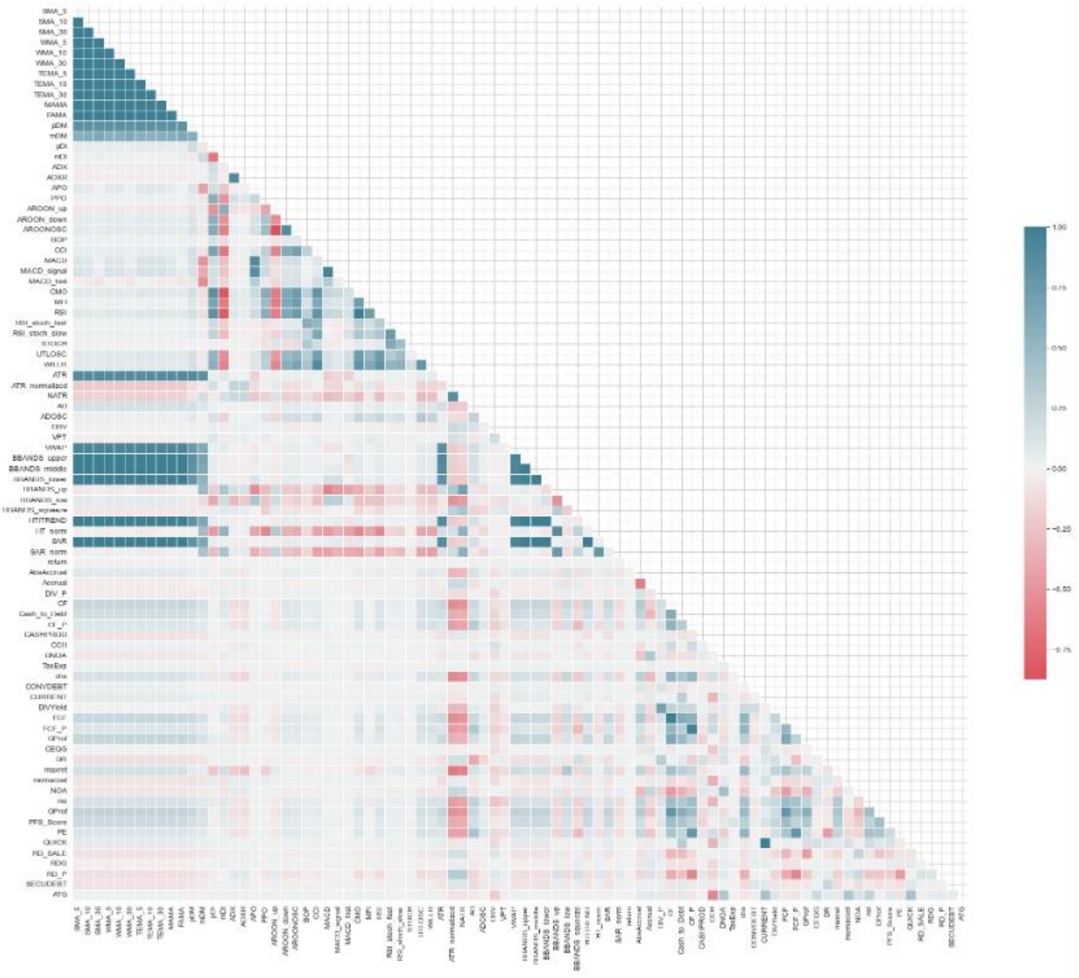


Figure 1: Factor Correlations

We can briefly see it through this picture:

Between moving average factors (SMA_5, SMA_10, SMA_30): They are highly positively correlated with each other.

P/E and P/B : These two financial indicators have a strong positive correlation (close to deep blue), as they both reflect the dimensions of company valuation

ATR_normalized (volatility) and BBANDS_width: They are highly positively correlated and appear in dark blue, indicating that the width of Bollinger bands increases when volatility is high

P/E and ROE: These are low or nearly uncorrelated (the color is nearly white), reflecting the fact that P/E and ROE are not directly related to each other.

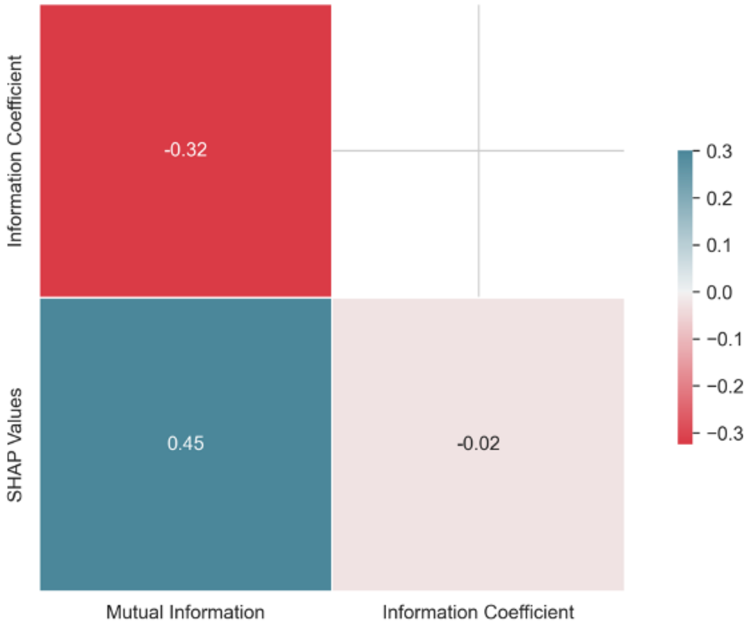


Figure 2: Rank Correlation of Feature Metrics

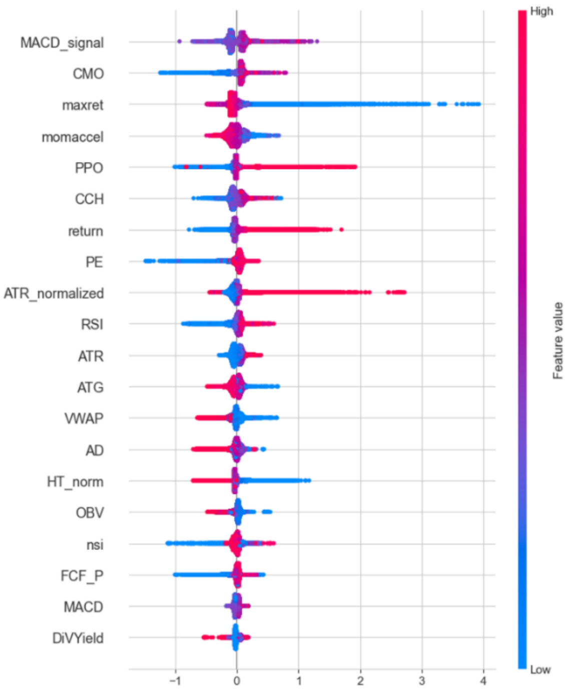


Figure 3: SHAP value

5.1.2 Rank correlation

The correlation between SHAP value and Mutual Information is high: the correlation coefficient is 0.45, indicating that there is some consistency between the results of these two metrics on feature evaluation, which is suitable for joint use.

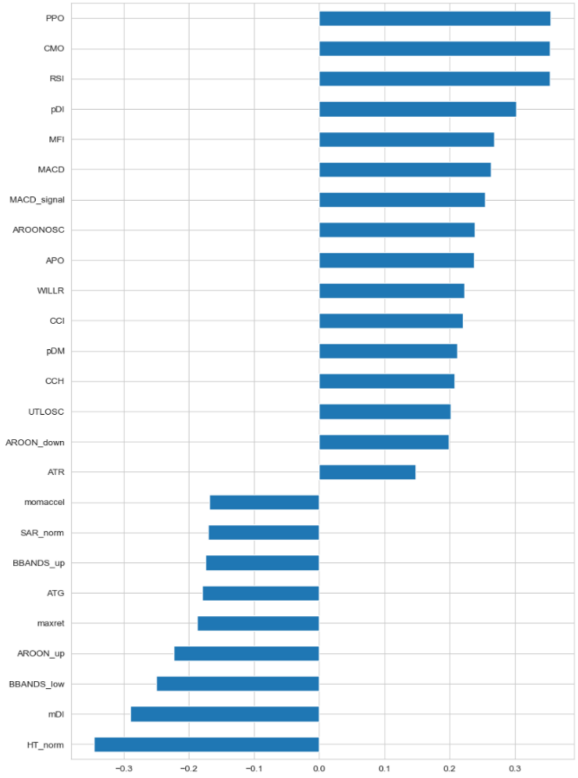
The correlation between Mutual Information and Information Coefficient is negative: the correlation coefficient is -0.32, suggesting that the two metrics are somewhat different in terms of feature evaluation and may be suitable for separate use in different scenarios

There is almost no correlation between the SHAP values and the information coefficients: the correlation coefficient is -0.02, indicating that they evaluate the importance of the features in completely different ways and can be used complementarity.

5.1.3 SHAP value

MACD_signal, CMO, maxret and momaccel are the main features affecting the predicted output of the model, and their SHAP values are widely distributed, indicating that these features are of high importance to the model.

Features such as MACD_signal and CMO have a large positive push on the model at high values (red) (positive SHAP values).



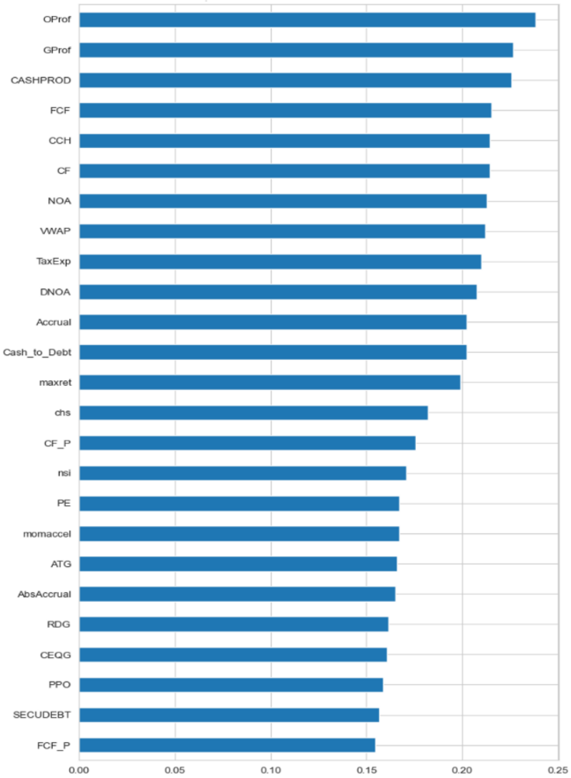


Figure 4: Factors correlation with alpha **Figure 5: Factors mutual information**

Positively correlated factors: e.g. PPO (Percentage Price Oscillator), CMO (Chande Momentum Oscillator), RSI (Relative Strength Index), PDI (Positive Directional Indicator). have a strong positive correlation to returns. These factors may contribute more to return forecasting or alpha generation. Used to construct long signals.

Negative correlation factors: such as HT_norm (Hilbert Transform Norm), mDI (Minus Directional Indicator), and BBANDS_low (Lower Bollinger Band) are significantly negatively correlated with yields.

These negatively correlated factors may have an inverse effect on yield forecasts and can be used to construct short signals.

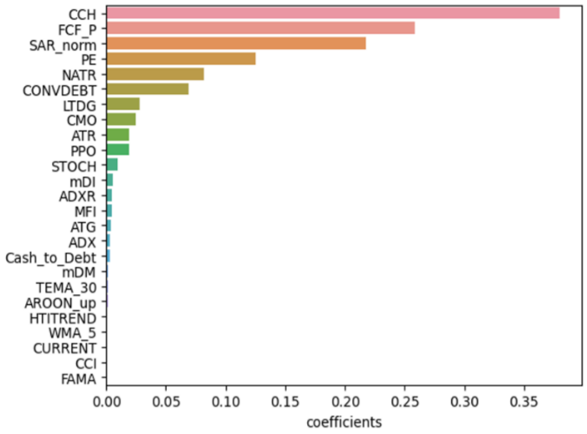
Factors with high information contribution: The top factors such as OProf (Operating Profitability), GProf (Gross Profitability), CASHPROD, FCF (Free Cash Flow), etc. have high mutual information values, which indicates that witheir relationship with the return is strong and potentially complex. These factors correlate with a company's financial health and profitability, and may provide a strong predictive basis for medium- to long-term investment strategies.

Other financial factors such as CF (Cash Flow), NOA (Net Operating Assets), and Cash_to_Debt also rank high, reflecting the significant impact of cash flow management and debt levels on returns.

Trading factors such as VWAP (Volume Weighted Average Price), momaccel (Momentum Acceleration) show that technical indicators still contribute to return forecasting.

5.2 The relationship between alpha and factors in 4 models

5.2.1 coefficients of linear models



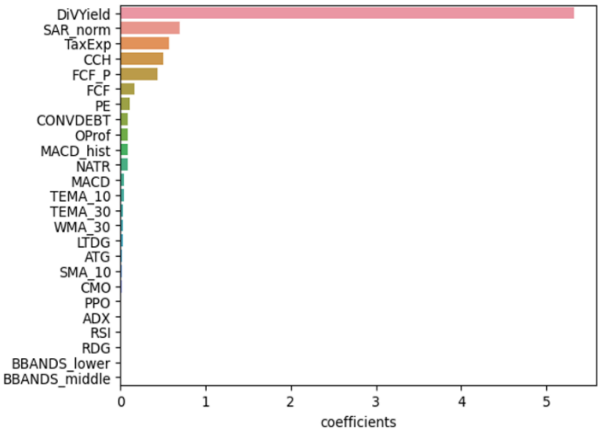
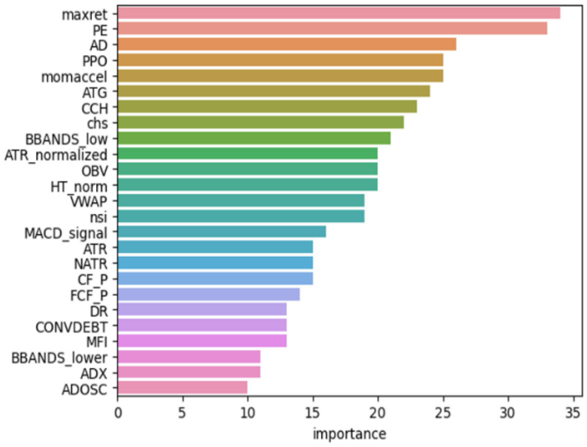


Figure 6: Coefficients of Elastic Net **Figure 7: Coefficients of Elastic Net**

Under the Elastic Net model, the coefficient values of CCH, FCF_P and SAR_norm are large, indicating that these three factors have a significant effect on alpha generation. Under the Ridge model, the coefficient values of DiVYield, SAR_norm, and TaxExp are larger, indicating that these three factors have a significant effect on alpha generation.

5.2.2 feature importance of boosting models



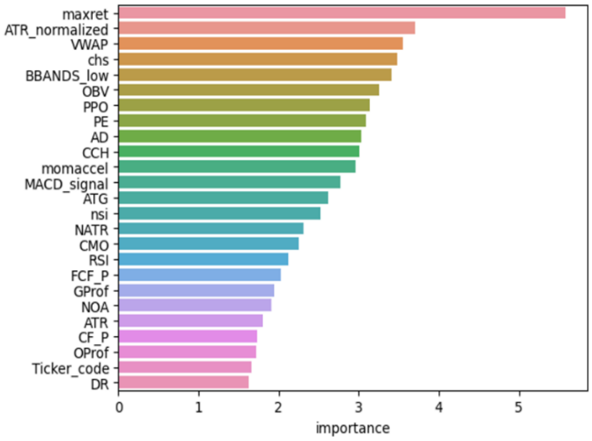


Figure 8: Feature importance of Lightgbm **Figure 9: Feature importance of Catboost**

Under the Light GBM model, the importance values of maxret, PE and AD are larger, indicating that these three factors have a significant effect on alpha generation. Under the Catboot model, the importance values of maxret, ATR_normalized, and VWAP are larger, indicating that these three factors have a significant effect on alpha generation.

Across the four models, we find that the factors CCH, FCF_P, PE, SAR_norm, BBANDS_low, ATR, and ADX recur, suggesting that they have a high degree of stability and significance in explaining the target variable (e.g., alpha or yield). These factors encompass both fundamental company metrics (e.g., financial health and valuation levels) and key market signals from technical analysis.

In summary, companies should prioritize these factors when constructing an investment strategy or optimizing company management.