A Report of Group Project in COMP5152

Quantitative Financial Trend Forecasting Based on Qlib

MA Zhiyuan 24039378G

ZHOU Letian 24068286G

April 4th, 2025

Table of Contents

Introduction

Quantitative finance is the core technology in the field of modern investment. It analyzes historical data and uses algorithmic models to predict market trends, thereby assisting institutions and individuals in making scientific decisions.

The main purpose is to use the open-source quantitative financial tool Qlib to develop a program to predict stock return trends, compare the performance of different models, including LightGBM, XGBoost and MASTER, to explore the trend prediction of financial data, and give investment advice. Through the practice of quantitative financial technology, you can further understand data analysis and the field of quantitative finance.

Data Source and Transformation

Data Source

Qlib has a variety of built-in data sets and can also handle user-defined data. Some of the built-in data sets include high-frequency and low-frequency data sets for the US and Chinese markets. Examples of these data sets are the **Alpha360** and **Alpha158** datasets, which are used for factor research and model training. There is also high-frequency data for short-term market analysis, such as 1-minute data. These data sets have been adjusted, for example, the price is standardized to 1 on the first trading day, and the original value can be fixed through $factor.

Also, users can convert their own CSV data to the .bin binary file format supported by Qlib. They can use tools such as `scripts/dump_bin.py` for conversion. The data must contain stock names (such as SH600000.csv) and date columns in a specific format.

Data Structure

Qlib uses a special flat file database (flat file database) design to convert data into a compact tree-structured binary format. The data is organized by financial instrument (e.g., stock) and attributes (e.g., opening price, closing price, etc.) and stored in different files. The timeline index is stored separately in a file named "calendar.txt". The first 4 bytes of each data file are set to the timeline index value, indicating the starting timestamp of the series of data. All data are arranged in chronological order. This design makes data updates (by appending) very efficient.

Data Transformation

Data Acquisition

The study uses constituent stock data from the CSI 300 Index (China) as the base dataset. The GetData() function provided by Qlib is utilized to access raw data from the database. Additionally, the Qlib API and the Alpha158 toolkit are used to extract 158 predefined factors, providing rich analytical dimensions for the study.

Data Conversion

The extracted raw data, originally in CSV format, is converted into .bin format using Qlib. This high-performance, compressed storage format is optimized for scientific computing and accelerates efficient data processing.

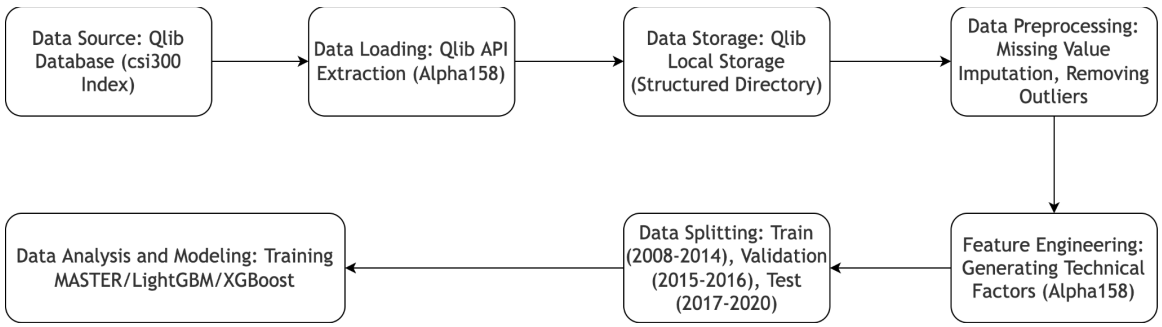


Figure 1 Flow of Data Processing

Data Alignment

Qlib uses infer_processors and learn_processors to preprocess the data, align features and labels, and ensure consistency. By filtering invalid data, normalizing and filling missing values, and sampling time periods in components such as TSDatasetH and TSDataSampler, features and labels are accurately aligned by time index, providing high-quality data for model training and evaluation.

Data Cleaning

Before analysis, the data undergoes cleaning, including handling missing values, removing outliers, and filling empty data points where necessary, ensuring that the dataset is of high quality.

Normalization

For normalization, all prices, including open, close, high, and low, are adjusted to start from a baseline value of 1 by dividing them by the first day's closing price. Additionally, prices are further adjusted for stock splits and

dividends using the factor field, where cumulative factors over time are applied to scale the prices accordingly. Volumes are also normalized inversely to price changes to maintain the original trading value, with the adjusted volume calculated as the raw volume multiplied by the ratio of the first day's close to the factor.

The normalized data are then stored in binary files, such as close.bin, which can be efficiently read using *numpy.fromfile()*. This format ensures that the data are compact and readily accessible for further processing.

Feature Engineering

In the data preprocessing phase, the Alpha158 and Alpha360 libraries provide extensive support. These libraries integrate a wide range of built-in expressions to calculate technical indicators. For example, you can calculate the 20-day momentum of a stock's closing price using the expression $close/Ref($close, 20). These libraries offer a rich set of predefined factors, including 158 factors in Alpha158 and 360 factors in Alpha360. These factors cover various aspects of technical analysis and financial modeling. You can also calculate custom features. This lets researchers make the library fit their specific research needs. This flexibility allows the incorporation of specialized knowledge and new ideas into the factor generation process, enhancing the adaptability and effectiveness of the platform for different quantitative investment strategies.

		KMID	KLEN	KMID2	KUP	KUP2	KLOW	KLOW2	KSFT	KSFT2	OPEN0	...
datetime	instrument											
2015-01-05	SH600000	0.008155	0.043286	0.188392	0.011286	0.260726	0.023845	0.550882	0.020714	0.478548	0.991911	...
	SH600008	0.044041	0.082901	0.531251	0.033679	0.406249	0.005181	0.062499	0.015544	0.187501	0.957816	...
	SH600009	0.006373	0.044608	0.142857	0.018627	0.417583	0.019608	0.439560	0.007353	0.164834	0.993668	...
	SH600010	0.063884	0.066340	0.962978	0.002456	0.037022	0.000000	0.000000	0.061428	0.925957	0.939952	...
	SH600011	0.027442	0.050494	0.543479	0.010977	0.217391	0.012075	0.239130	0.028540	0.565217	0.973291	...

Figure 2 Screenshot of Data Structure for Example

Dataset Splitting

The data is divided into three parts in chronological order: data from 2008 to 2014 is used for training, data from 2015 to 2016 is used for verification, and data from 2017 to 2020 is used for testing. This forward and backward segmentation method is in line with the pertinence of the financial market and effectively prevents excessive behavior problems caused by the leakage of "future data".

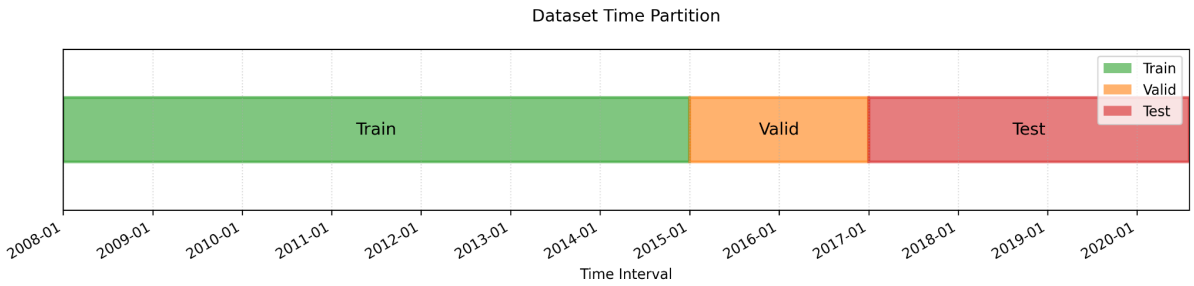


Figure 3 Segmentation for Training, Validing and Testing

Model Building

In the modeling phase, algorithms such as the MASTER model, LightGBM, and XGBoost are employed for model training. The models' performances are compared, and the best-performing model is selected for detailed quantitative analysis.

Data Visualization

There are two key visualizations that present the performance and behavior of the SH000300 stock over a certain period as an example from about more than 200 assets in CSI300.

The first diagram highlights the Stock Ups and Downs Trend Chart, capturing the daily price fluctuations of the stock to provide insights into its overall volatility and directional movement. This allows us to observe the patterns of stock price increases and decreases over time, identifying potential periods of sharp price changes or stability.

Figure 4 An Example of SH000300 Stock Trend

This chart shows the trend of a financial asset SH000300 (daily yield) (e.g., a stock, portfolio or index) over the period from 2008 to 2020. The daily return in the graph is the percentage change in the daily price of the asset, reflecting short-term volatility and the volatile nature of the market. Positive and negative values appear in the chart because the daily yield is calculated to reflect the direction and magnitude of the daily change in the asset's price.

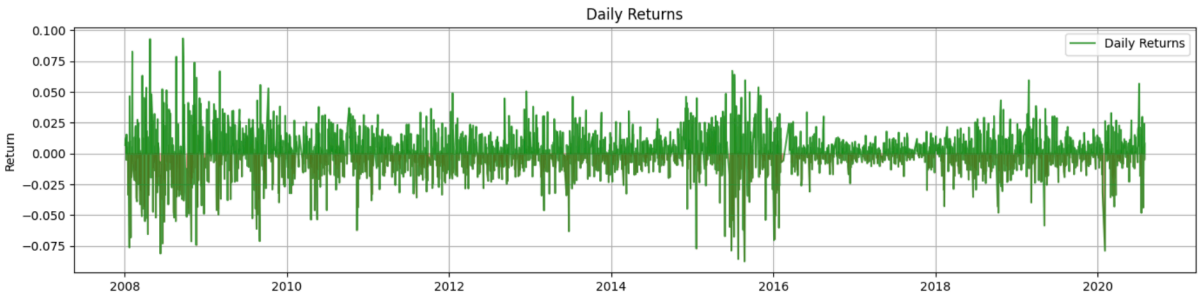


Figure 5 Diagram of Daily Return Time Series

The daily yield is a percentage change value. It is calculated as the change in an asset's price between two trading days. If the asset's price is higher on the next day, the daily yield is positive. This means that the asset's value

increased, and the investor earned a profit. If the closing price of an asset on a given day is lower than the previous day, the calculated daily rate of return is negative. This means that the asset lost value on that day, and the investor lost money.

Analysis Steps and Trials

Data feature analysis and modelling

For different financial forecasting tasks, this project selected a variety of models for comparison:

Classic time series models:

LGBM[1]**:** LightGBM is a gradient-boosting framework that uses tree-based learning algorithms. It is designed to be distributed and efficient with the following advantages:

- Faster training speed and higher efficiency.
- Lower memory usage.
- Better accuracy.
- Support of parallel, distributed, and GPU learning.
- Capable of handling large-scale data.

However, the LGBM model can be prone to overfitting, especially with small datasets, and requires careful hyperparameter tuning to achieve optimal performance. While it handles large-scale data well, memory usage can still be significant for very high-dimensional datasets. Additionally, interpreting the model's decisions can be challenging due to its complexity.

Machine learning model:

XGBoost[2]: a novel sparsity-aware algorithm for sparse data and weighted quantile sketch for approximate tree learning. XGBoost model also introduces insights on cache access patterns, data compression and sharding to build a scalable tree boosting system.

Reinforcement learning model:

MASTER: a MArkert-Guided Stock TransformER, which models the momentary and cross-time stock correlation and leverages market information for automatic feature selection. Existing works based time-aligned stock correlations stemming are not able to handle the unpredictable cross-time correlations and judge the effectiveness of dynamic features. However, MASTER elegantly tackles the complex stock correlation by alternatively engaging in intra-stock and inter-stock information aggregation[3].

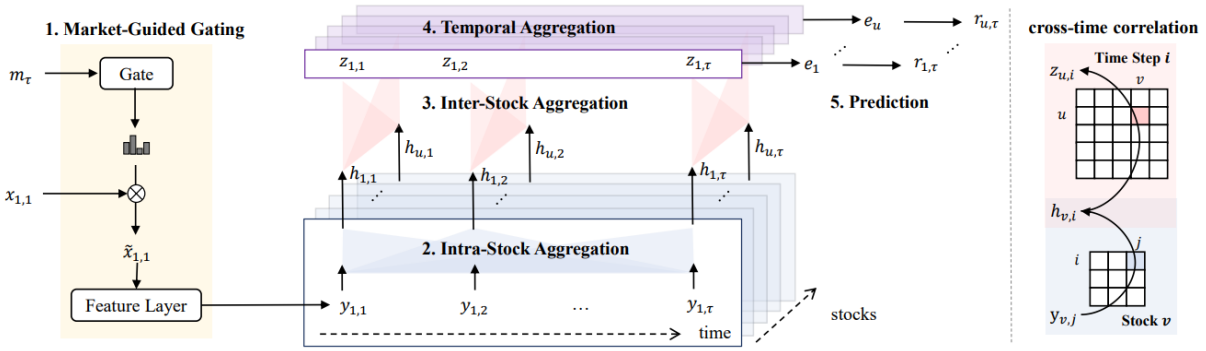


Figure 6 Diagram of MASTER Structure

The Workflow of the MASTER model is illustrated as follows:

(1) Market-Guided Gating. A vector representing the current market status $m\tau$ was constructed and leveraged to rescale feature vectors by a gating mechanism, achieving market-guided feature selection.

(2) Intra-Stock Aggregation. Within the sequence of each stock, at each time step, MASTER aggregates information from other time steps to generate a local embedding that preserves the temporal local details of the stock while collecting all important signals along the time axis. The local embedding *hu,t* will serve as relays and transport the collected signals to other stocks in subsequent modules.

(3) Inter-Stock Aggregation. Calculate the correlation between each stock with **attention mechanism**, and each stock further aggregates the local embedding of other stocks. The aggregated information *zu,t*, which was referred to as temporal embedding, contains not only the information of the momentarily correlated stocks at *t*, but also preserves the personal information of *u*.

(4) Temporal Aggregation. For each stock, the last temporal embedding queries from all historical temporal embedding produce a comprehensive stock embedding *eu*.

(5) Prediction. The comprehensive stock embedding is sent to prediction layers for label prediction.

Model training and validation

This section presents the training and validation processes for the three models—**MASTERModel**, **XGBModel**, and **LGBModel**—applied to the CSI 300 market. These models are tasked with analyzing stock market trends, generating predictions, and supporting portfolio optimization through carefully configured data pipelines.

The dataset covers the period from **2008-01-01** to **2020-08-01** and is split into training (2008-2014), validation (2015-2016), and testing (2017-2020) segments. Preprocessing, which is consistent across models, uses the Alpha158 handler. Techniques like robust Z-score normalization, filling missing values, cross-sectional rank

normalization, and handling missing labels ensure high data quality. The label is defined as the 5-day percentage return: Ref($close, -5) / Ref($close, -1) - 1. The processed data is fed into the DatasetH or MASTERTSDatasetH to ensure robust and segmented handling for modeling and evaluation.

The **MASTERModel** employs a deep learning architecture specifically designed for sequential time-series analysis. It uses 40 epochs for training, a learning rate of 0.000008, and early stopping to avoid overfitting. The training phase uses data from 2008 to 2014, while validation on the 2015-2016 segment evaluates metrics like IC and precision. During testing on the 2017-2020 period, realistic backtesting is conducted using a strategy that selects the top 30 signals and drops the lowest-performing ones, with results benchmarked against the CSI 300 index to assess financial performance.

The **XGBModel** leverages XGBoost's gradient boosting algorithm for regression tasks, configured with 647 boosting iterations, a maximum tree depth of 8, and feature and data sampling ratios to prevent overfitting. The training process aligns with the 2008-2014 period, while the validation phase optimizes model parameters using RMSE as the evaluation metric. Backtesting during the testing phase assesses signal quality through the SignalRecord and SigAnaRecord, while portfolio returns are analyzed with the PortAnaRecord class for practical strategies. This setup highlights XGBoost's capability for precise tabular regression tasks.

The **LGBModel** uses LightGBM for efficient gradient boosting on large datasets. The model is configured with a learning rate of 0.0421, a maximum depth of 8, and 210 leaves to capture complex feature interactions. Regularization through L1 and L2 penalties prevents overfitting, and multithreading enhances computational efficiency. Training data from 2008-2014 is followed by validation on the 2015-2016 segment to fine-tune parameters. Testing on 2017-2020 evaluates model generalization, and backtesting ensures alignment with real-world use cases, showcasing the model's ability to provide practical portfolio management insights.

In summary, the three models—MASTERModel, XGBModel, and LGBModel—are rigorously trained and validated using robust preprocessing, segmented datasets, and distinct evaluation strategies. While the MASTERModel's deep learning approach excels in sequential data analysis, the tree-based frameworks of XGBoost and LightGBM provide efficient solutions for tabular data regression. The comprehensive pipeline ensures robust predictive performance and reliable portfolio optimization in the CSI 300 market.

Model Evaluation

Those three charts visualize the performance of three models—**LGBM**, **Master**, and **XGBoost**—across a set of metrics. The first chart, a **grouped bar chart**, shows the absolute values of the performance metrics for easier direct comparison among the models. The second chart, a **radar chart**, offers a comprehensive view of trends and model differences by representing the metrics as polygons.

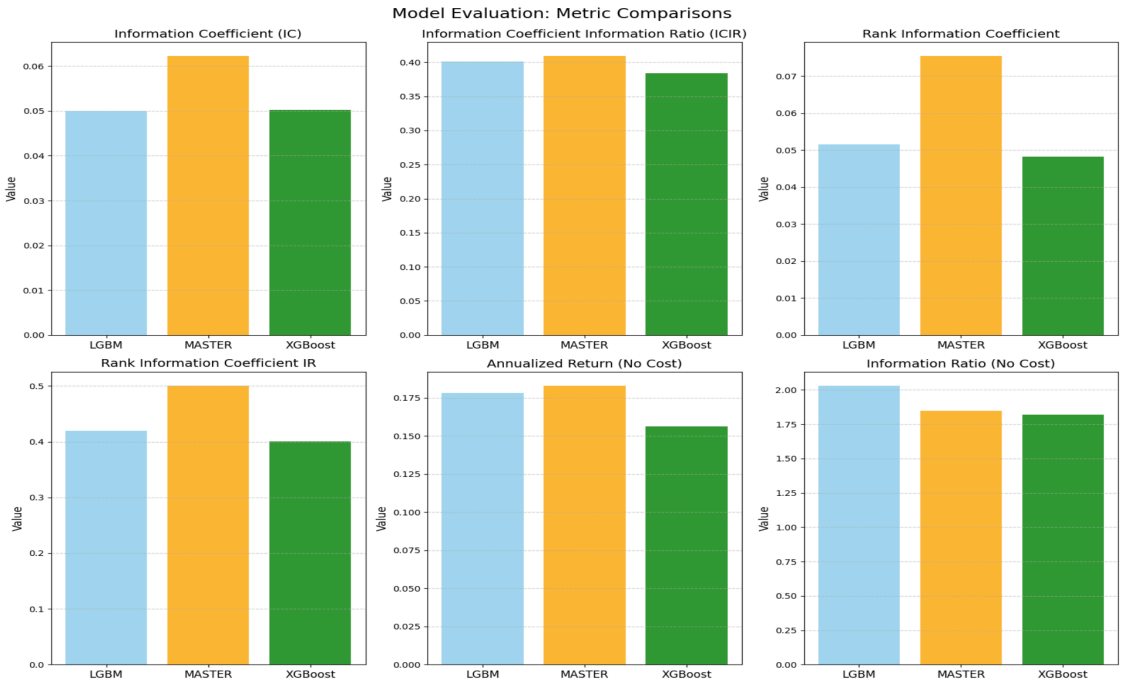


Figure 7 Bar Charts for Comparison by 6 Attributes

In the first chart, specific differences in performance are more apparent. For the **IC (Information Coefficient)** metric, Master_result achieves the highest value of 0.0622, while XGBoost_result and LGBM_result closely follow with values of 0.0503 and 0.0499, respectively, showing only a marginal difference between the top and bottom performers. The **ICIR (Information Coefficient Information Ratio)** metric reflects comparable results for all models, with Master_result slightly leading at 0.4089, followed by LGBM_result at 0.4012 and XGBoost_result at 0.3842. For the **Rank IC**, Master_result outperforms others with a score of 0.0754, while LGBM_result (0.0515) edges out XGBoost_result (0.0482). Similarly, in the **Rank ICIR**, Master_result is the top performer with 0.5004, with LGBM_result and XGBoost_result trailing at 0.4196 and 0.4009, respectively. A slight divergence is observed in **ERWC_AR (Excess Return Weighted Correlation - Absolute Return)**, where Master_result leads with 0.1830, followed closely by LGBM_result (0.1781), while XGBoost_result lags slightly behind at 0.1565. However, **ERWC_IR (Excess Return Weighted Correlation - Information Ratio)** distinguishes LGBM_result as the dominant model, boasting an extraordinary score of 2.03, far surpassing Master_result (1.85) and XGBoost_result (1.82).

The radar chart complements these observations by visualizing the broader trends across metrics. LGBM_result stands out sharply in **ERWC_IR**, forming the most pronounced peak, underscoring its specialization in this metric. Master_result offers consistent and balanced performance across most metrics, excelling in **IC**, **Rank IC**, and **Rank ICIR** while maintaining competitive scores in others. XGBoost_result, in contrast, demonstrates relatively weaker performance, particularly in **Rank IC** and **Rank ICIR**, while slightly narrowing the gap in **ICIR** and **ERWC_AR**, where all three models cluster closely. This central grouping indicates that these metrics exhibit minor variations among the models.

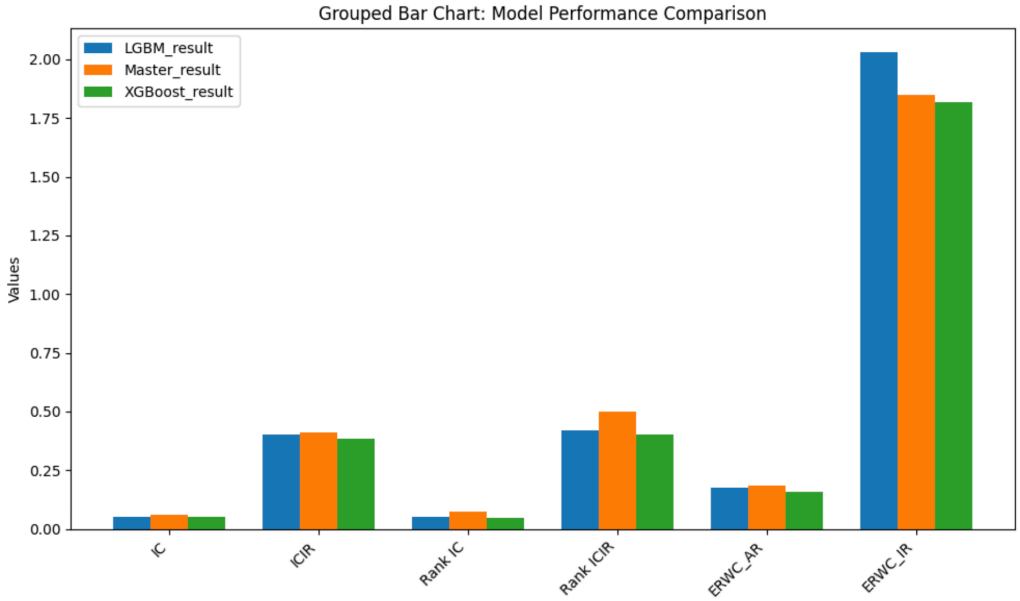


Figure 8 Bar Chart of Overall Comparison

In summary, Master_result emerges as the top performer overall, with leading scores in **IC**, **Rank IC**, and **Rank ICIR**, making it a balanced and reliable choice for general performance across diverse metrics. LGBM_result, on the other hand, is a specialist in **ERWC_IR**, where it dominates significantly, making it the best option for use cases where this metric is critical. XGBoost_result consistently lags behind the other two models and may require optimization to achieve competitive performance in key areas. These analyses support selecting models based on the importance of specific metrics to the task at hand.

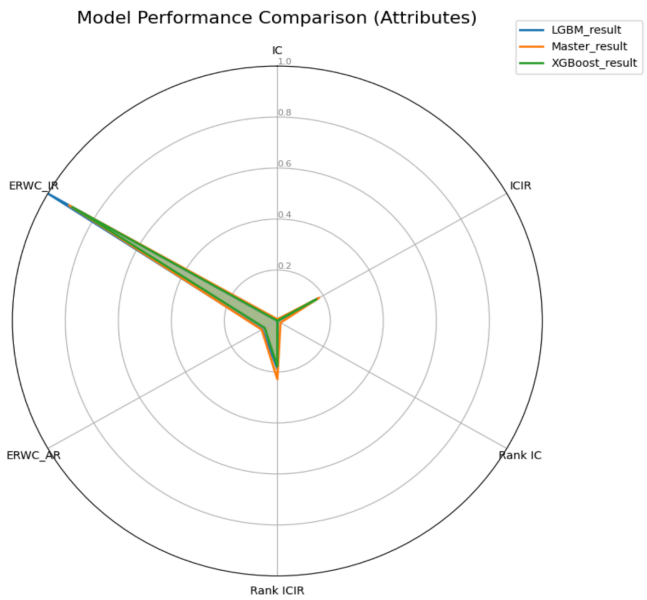


Figure 9 Radar Chart for Model Performance Comparison

Prediction and Conclusions

Based on the performance comparisons, the Master model demonstrates the most balanced and consistent performance across all key metrics. It outperforms the other models in critical areas such as Rank IC, Rank ICIR, and ERWC_AR. The LGBM_result model does very well in ERWC_IR, but its performance is not as consistent in the other areas, which makes it less suitable as a general model. The XGBoost_result model also shows weaker results in most metrics, meaning it needs to be improved to be competitive.

Overall, the Master model performs better, so it's been chosen to be the main model for future analyses. The Master model consistently produces reliable predictions and robust analytical results, making it the best choice for achieving consistent results across various performance attributes. In the future, we will focus on improving and using the Master model to learn more, using its strengths to make it easier to understand and more accurate.

Report

This chart comprehensively shows the long-term performance comparison of different investment strategies, which is a typical display of quantitative investment backtest results. The chart uses time as the horizontal axis, and is divided into seven sub-chart panels vertically. Each panel presents the strategy performance from a different dimension, allowing investors to comprehensively evaluate the pros and cons of various investment methods.

Cumulative Return Comparison

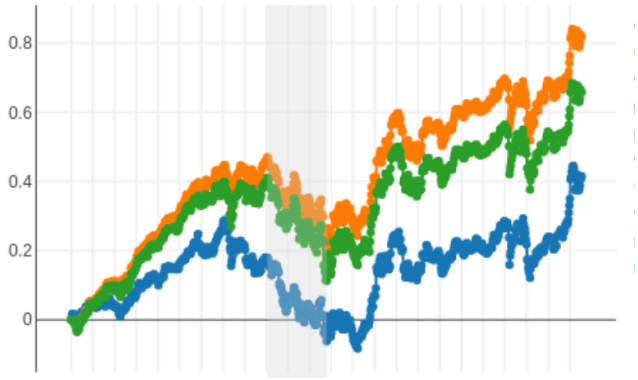


Figure 10 Diagram of Cumulative Return Comparison

The top panel has three lines that show how the returns have added up over time. The blue line shows the performance of the benchmark investment (cum bench), which represents the return rate of the market index or other reference target. It had some ups and downs during the period, and it even dropped to almost zero at times. However, it showed a clear upward trend at the end, eventually reaching a return of about 40%. "Benchmark" is usually used as a reference to measure the success or failure of active strategies. Some common benchmarks include the S&P 500 Index and the CSI 300.

The orange line shows the total return of the strategy without considering transaction costs (cum return wo cost). This strategy did very well, especially in the middle and late stages, and finally reached a return of about 80%, becoming the best performance of all strategies. "Not considering transaction costs" means that this is the pure strategy return under ideal circumstances, without deducting various costs such as transaction commissions, stamp duties, slippage, etc.

The green line shows the return after considering these costs. It's similar to the orange line, but the yield is slightly lower, about 65%. This shows the impact of transaction costs on long-term investment performance. This line is closer to the return that can be obtained when investing in reality.

Risk Metrics Dashboard

The second panel uses a red area chart to show the return without the maximum drawdown (MDD) adjustment (return w/o MDD). This chart shows a negative area and clearly shows the down cycle in the investment process. "Maximum drawdown" refers to the most loss a portfolio can have from its highest point to its lowest point. It is an important way to measure investment risk. This indicator has fluctuated significantly at multiple time points, with the worst drop of about 30%, revealing the short-term risks faced by the strategy.

The purple area chart immediately below shows the return after MDD adjustment[4] (return w cost MDD), which is similar to the red area pattern, but represents the risk-adjusted return performance. This is a better way to evaluate a strategy because high returns usually mean high risk.

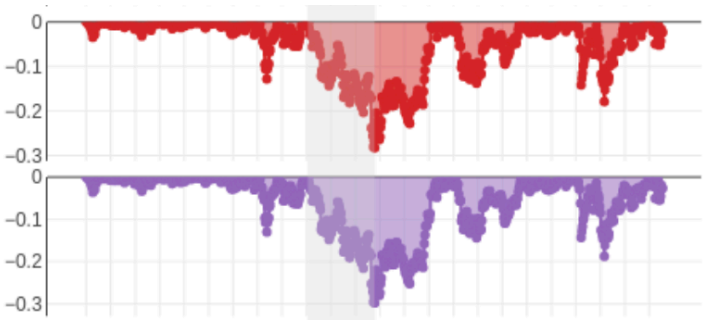


Figure 11 Diagram of Risk Metrics

Excess Return Panel

The lower center panel uses brown and pink lines to track the total extra returns without and with costs (cum ex return with cost and cum ex return with cost). Excess Return is the difference between the strategy return and the benchmark return. A positive value indicates that the strategy outperforms the benchmark. The brown line shows the extra return from the strategy compared to the benchmark. This extra return is most noticeable when transaction costs are not considered, and it ends up being about 40%. The pink line shows the excess return after transaction costs are considered. Although it is lower than the brown line, it still reaches an excess return of about 25-30%, proving that even considering costs, the strategy still significantly outperforms the benchmark.

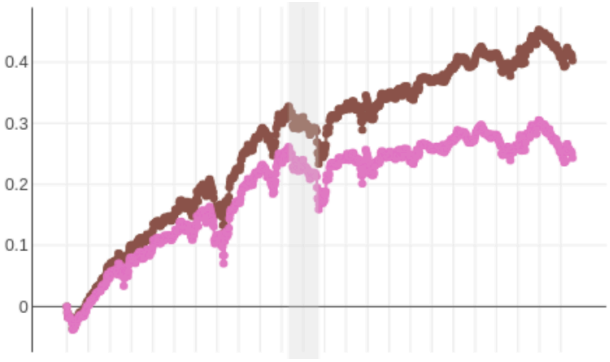


Figure 12 Diagram for Excess Return Panel

Turnover and Risk Adjustment Panel

Several panels at the bottom of the chart provide additional information. The gray line may represent the portfolio turnover rate, which, except for a significant high point in the early stage, has remained at a low level most of the time, indicating that the strategy is relatively stable and does not need frequent adjustments. The turnover rate reflects the frequency of portfolio adjustments. A high turnover rate usually means higher transaction costs and more aggressive trading strategies.

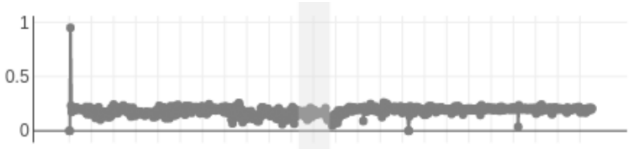


Figure 13 Diagram for Turnover

The yellow and cyan lines at the bottom show different excess return indicators (cum ex return w cost MDD and cum ex return wo cost MDD) including MDD adjustments. Both fluctuate in the negative area, revealing the challenges faced by the risk-adjusted strategy. This shows that although the absolute return and excess return of the strategy are good, its performance may not be as good as it seems after considering risk factors.

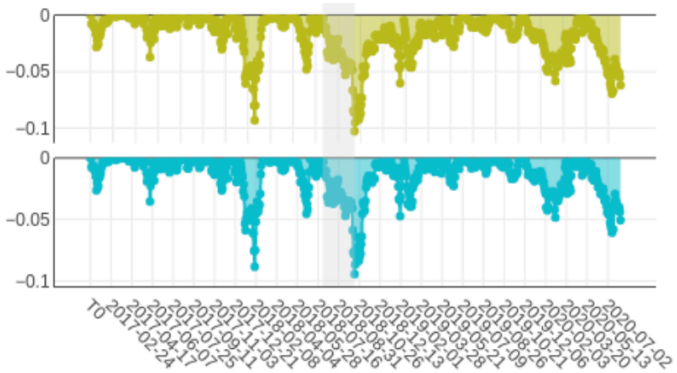


Figure 14 Diagram for Risk Adjustment Panel

Overall

There is a clear gray vertical area in the whole chart. This area could be a sign of important market events or times of high volatility (like financial crises or epidemic shocks). Almost all strategies have experienced significant fluctuations during this period. This makes it an important time to evaluate the strength of strategies.

These charts are very important for quantitative investment and strategy backtesting. They can help investors evaluate the performance of different strategies in different market conditions. By looking at many different factors at once, such as returns (both the total amount and how much they change over time), risks (the biggest losses possible), costs (how much it costs to buy and sell), and how easy it is to buy and sell (how often the price changes), investors can make better decisions and avoid getting caught up in trying to get high returns by ignoring possible risks.

Risk

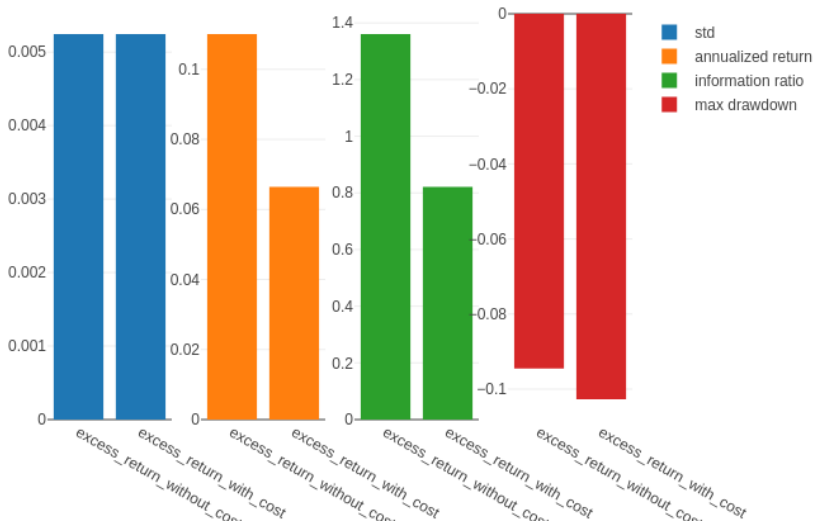


Figure 15 Diagram for MASTER's Performance

In such a figure, a comparative analysis of MASTER's performance is generated under two scenarios: "with cost" and "without cost," which likely refer to the inclusion or exclusion of transaction costs in the investment strategy.

The chart comprehensively illustrates the advancement of the MASTER model in metrics below:

- Standard Deviation (std): The chart shows that the strategy without cost has a slightly higher standard deviation compared to the strategy with cost. This suggests that including transaction costs may lead to slightly more stable returns, possibly due to reduced trading frequency.
- Annualized Return: The strategy with cost exhibits a higher annualized return. Indicating that despite the inclusion of costs, the strategy is **more effective** at generating returns, likely due to more prudent trading decisions that avoid unnecessary costs.
- Information Ratio: The strategy with cost has a higher information ratio, demonstrating superior risk-adjusted performance. This suggests that the cost-aware strategy is better at **maximizing return**s relative to the risk taken.
- Maximum Drawdown: The strategy with cost also shows a lower maximum drawdown, indicating a **more robust performance** during adverse market conditions. This is a significant advantage as it implies the strategy is better equipped to handle losses.

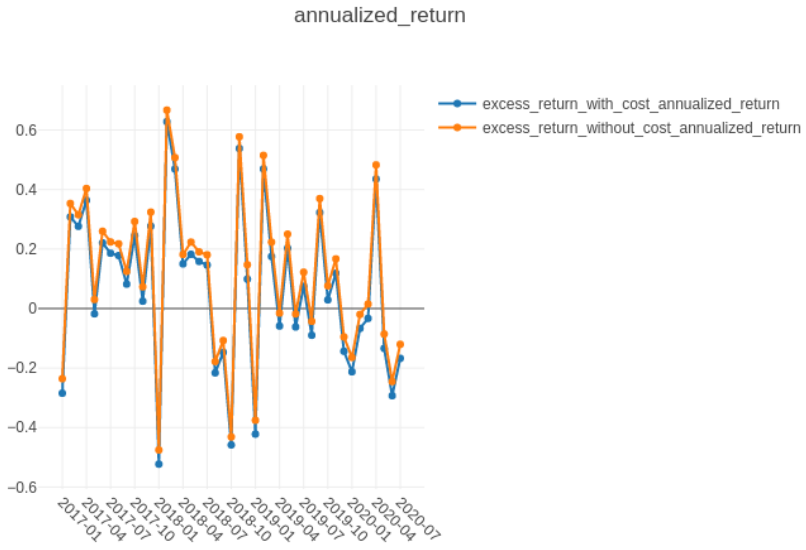


Figure 16 Diagram for Annualized Return

Content: This chart shows the annualized return of the investment strategy over time.

Interpretation: The strategy without cost generally shows higher peaks, indicating periods of higher returns. However, the strategy with cost exhibits more consistent performance, suggesting that while transaction costs reduce the potential for high returns, they also contribute to more stable performance.

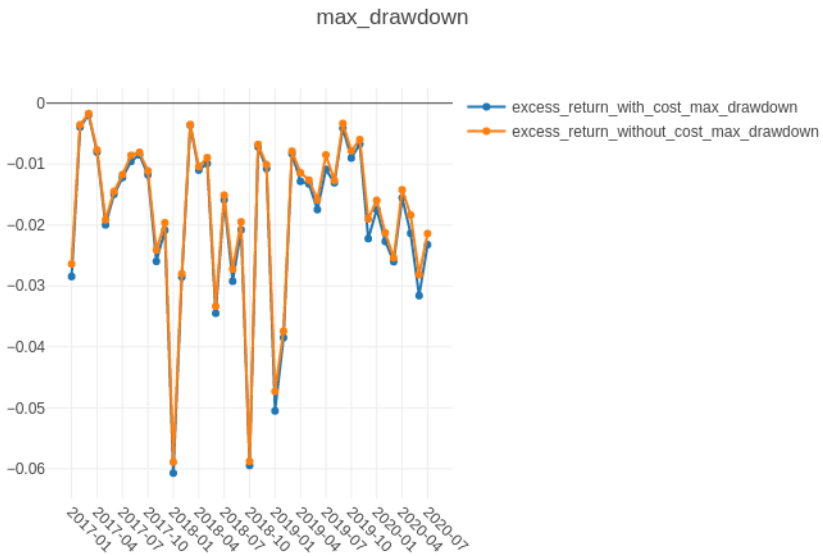


Figure 17 Diagram for Max Drawdown

Content: This chart illustrates the maximum drawdown, which is the maximum observed loss from a peak to a trough of a portfolio, before a new peak is achieved.

Interpretation: The strategy with cost tends to have a lower maximum drawdown, indicating that including transaction costs in the strategy can lead to less severe losses during market downturns, enhancing risk management.

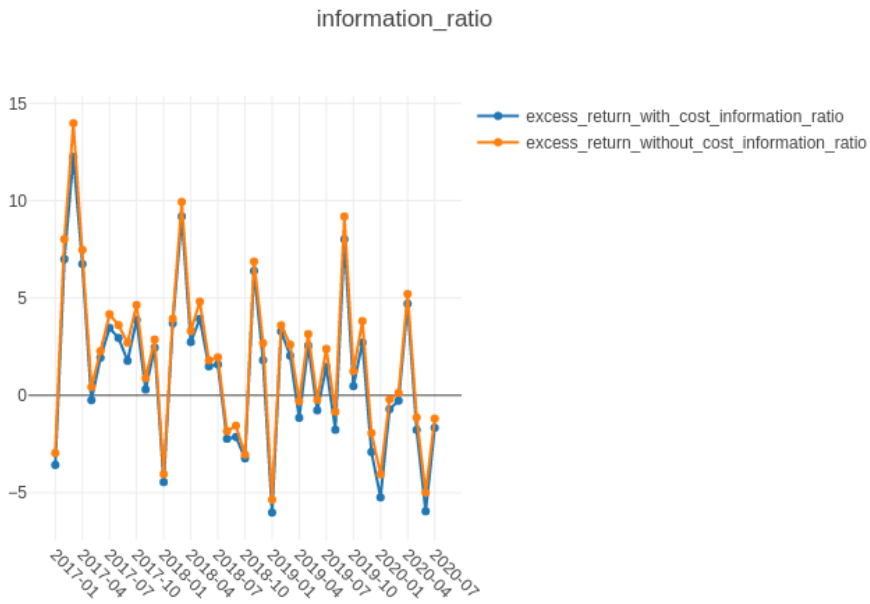
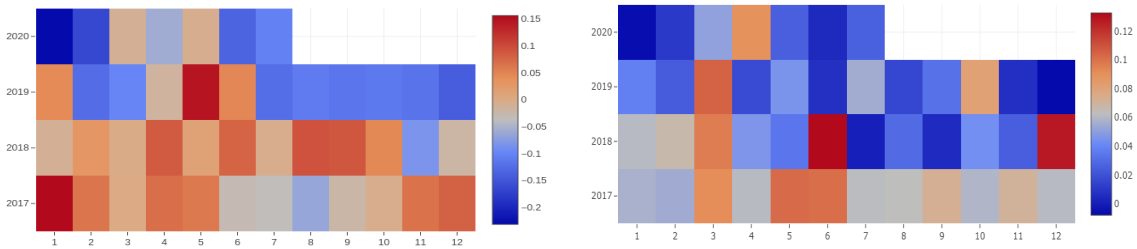


Figure 18 Diagram for Information Ratio

Content:

- The first chart displays the information ratio, which measures the excess return (return over the benchmark) per unit of deviation in excess return of the MASTER model and baseline model.
- The second set of figures compare the Monthly IC of MASTER and XGBoost, the MASTER model in the time period of the test dataset.



- *Figure 19 Comparison Heatmap for MASTER and XGBoost*

Interpretation: The strategy with cost shows a higher information ratio in the comparison of baseline and other models(e.g. XGBoost). Suggesting that it offers a better risk-adjusted return. This indicates that the cost-aware strategy is more efficient at generating excess returns relative to the risk taken.

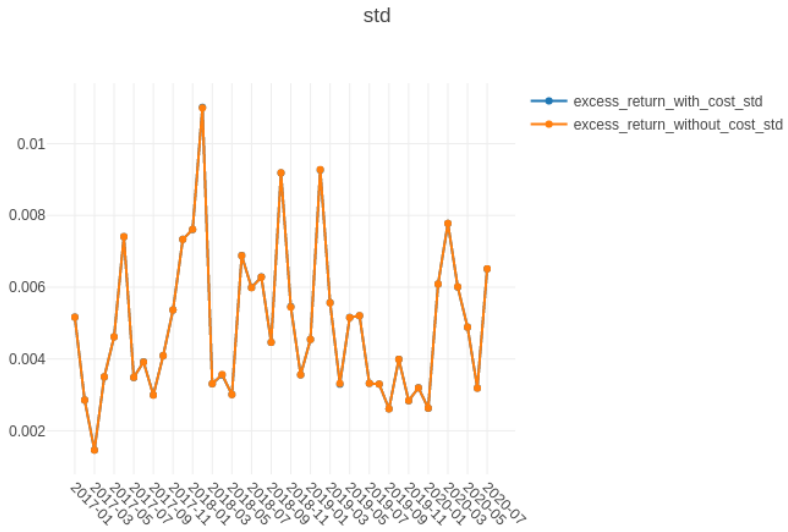


Figure 20 Diagram for Standard Deviation of Excess Returns

Content: This chart represents the standard deviation of excess returns, indicating the volatility of the strategy's returns.

Interpretation: The standard deviation is the same for both strategies, confirming that transaction costs do not influence the volatility of returns. This suggests that the model maintains a consistent level of risk regardless of the cost considerations

Score IC

This chart shows how IC scores have changed over time from 2017 to 2020. The chart is called "Score IC," which tells us that it shows information coefficient-related indicators. The chart has two curves: the blue line represents "ic" and the orange line represents "rank_ic." Both are clearly marked in the legend.

Overall, the two lines show very similar patterns, almost completely overlapping. This indicates that the performance of traditional IC and ranking IC on this particular factor or strategy is highly correlated. The IC score shows significant ups and downs over the whole period, with no clear upward or downward trend, but rather random fluctuations around zero. The vertical axis ranges roughly from -0.4 to 0.4, and most data points are concentrated between -0.2 and 0.2.

In early 2017, the IC value peaked around 0.4, suggesting strong predictive power during that time. Throughout 2017, IC remained high but volatile. In 2018, the fluctuations continued but weakened overall, with alternating positive and negative values and several strong positive peaks mid-year. By 2019, IC variability persisted but with no extreme values. In 2020, the IC trend shifted downward, showing more negative values, especially in the latter half, with troughs near -0.4.

The IC value[5], a key measure of a factor's ability to predict future returns, alternates between positive (good prediction) and negative (poor prediction) values in the chart. Most data lies between -0.2 and 0.2, indicating moderate overall predictive power of the factor.

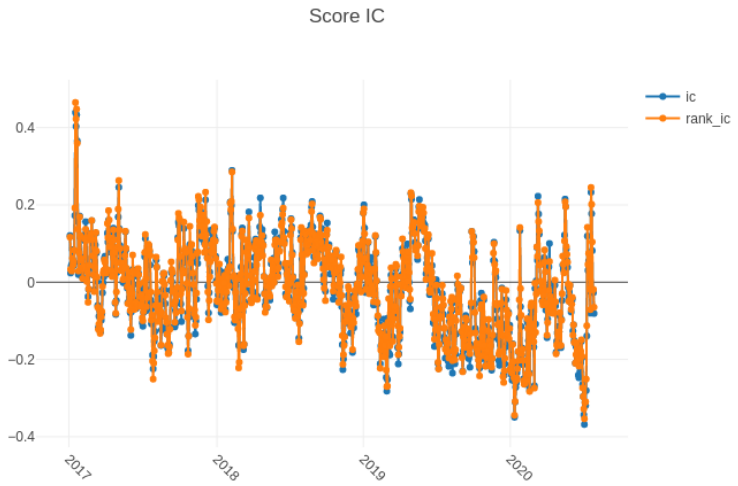


Figure 21 Diagram of Score IC

The overlap of traditional IC and ranked IC is rare, and it may indicate that the data is normally distributed or that the rankings are highly consistent with the original values. While there is no clear pattern over time, the ability to predict the future got weaker as time went on. This probably happened because the market learned to adapt to the factor signal. This is a common thing that happens in the investment world.

Performance

This set of charts comprehensively displays the analysis results of a quantitative investment strategy from multiple dimensions, including key indicators such as cumulative returns, strategy distribution characteristics, information coefficient performance and autocorrelation, providing us with a complete perspective for in-depth understanding of the performance of the strategy.

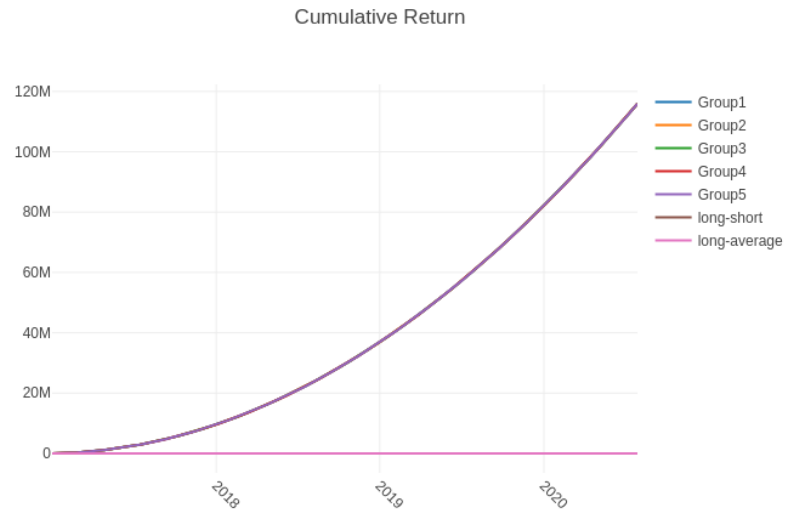


Figure 22 Diagram for Cumulative Return

The first chart shows the total returns from 2017 to 2020. The purple "long-short" strategy shows strong growth, especially after 2019, reaching almost 120 million by the end of 2020. The pink "long-average" strategy did not do as well, with only small returns. The other strategies were not as significant or did not show any significant results. This shows that the long-short strategy performed better.

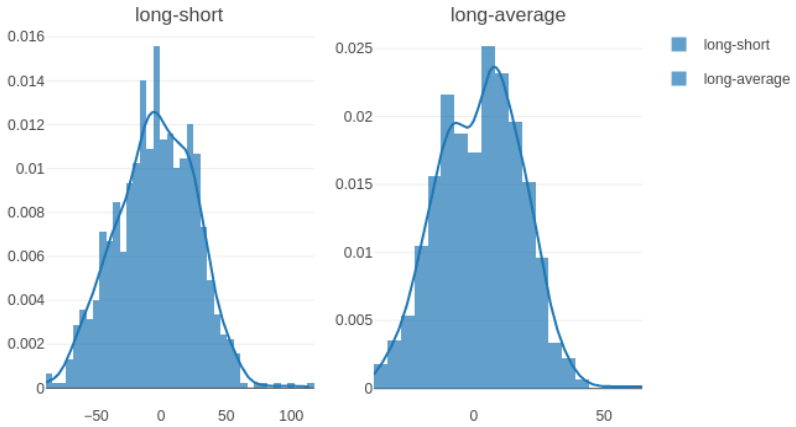


Figure 23 Diagram for Long-short and Long-average Return

The second chart shows the returns for each strategy. The "long-short" strategy has a normal distribution that is almost symmetrical, but slightly skewed to the right (from -75 to +75). This shows that there is moderate volatility and growth potential. The "long-average" strategy is more concentrated (0 to 50), with less volatility and limited returns, explaining its flat cumulative growth.

The third figure shows the time series of the Information Coefficient (IC) from January 2017 to July 2020. The blue and orange lines in the figure represent the IC and Rank IC, respectively, which almost completely overlap, indicating that the traditional correlation and ranking correlation are highly consistent in this strategy. The IC value fluctuates mainly between -0.4 and 0.4, mostly concentrated in the range of -0.2 to 0.2. It is worth noting

that in early 2017, the IC value showed a significant positive peak close to 0.4, and then fluctuated without obvious trend throughout the observation period. However, after entering 2020, the frequency of negative IC values seems to have increased, which may suggest that the effectiveness of the factor has weakened in the near future.

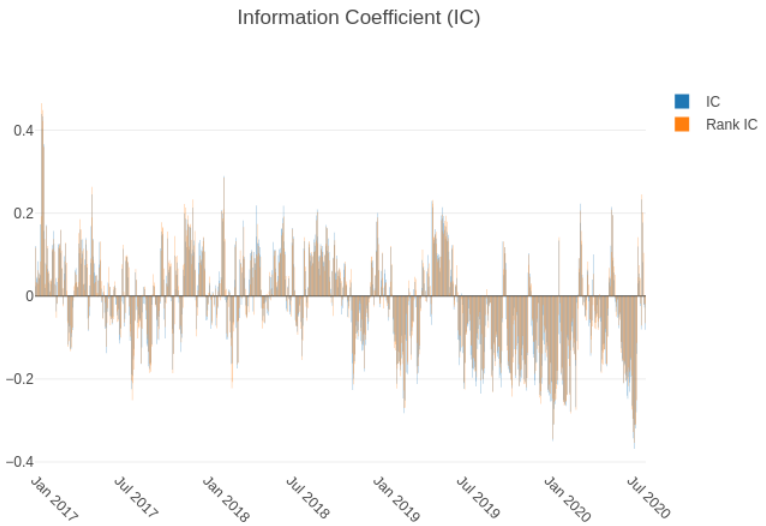


Figure 24 Diagram of Information Coefficient

The fourth figure shows the seasonal pattern of monthly IC values in the form of a heat map. From 2017 to 2020, the IC values of the 12 months of each year are represented by color depth, with red representing higher positive values and blue representing lower negative values. As can be seen in the figure, the early 2017 (especially January) is dark red, indicating that the IC value is the highest, which is consistent with the observations in the third figure. Overall, the IC values in 2017 and 2018 are generally positive (more red and orange), while the second half of 2019 and the first half of 2020 have more negative values (blue areas). This seasonal change provides important information and may reflect the impact of the market environment in a particular month on the effectiveness of the strategy.

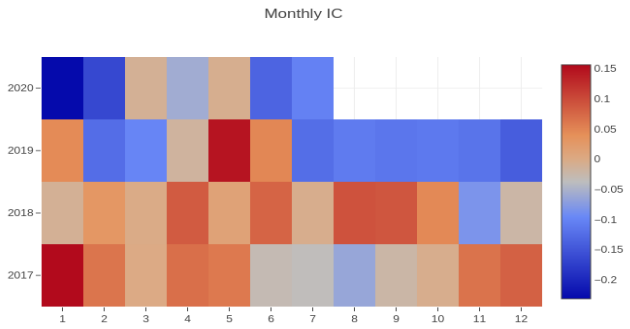


Figure 25 Heatmap for Monthly IC

The fifth figure shows a histogram of IC values on the left and a corresponding QQ plot on the right. The histogram shows that the IC values are generally approximately normally distributed, with the center slightly

biased toward the negative region, but with good overall symmetry. The QQ plot further verifies the distribution characteristics of the IC values, with most points arranged along the diagonal, indicating that the actual distribution is highly consistent with the theoretical normal distribution, with only slight deviations at both ends, which usually indicates that the tail of the distribution is slightly thicker than the standard normal distribution.

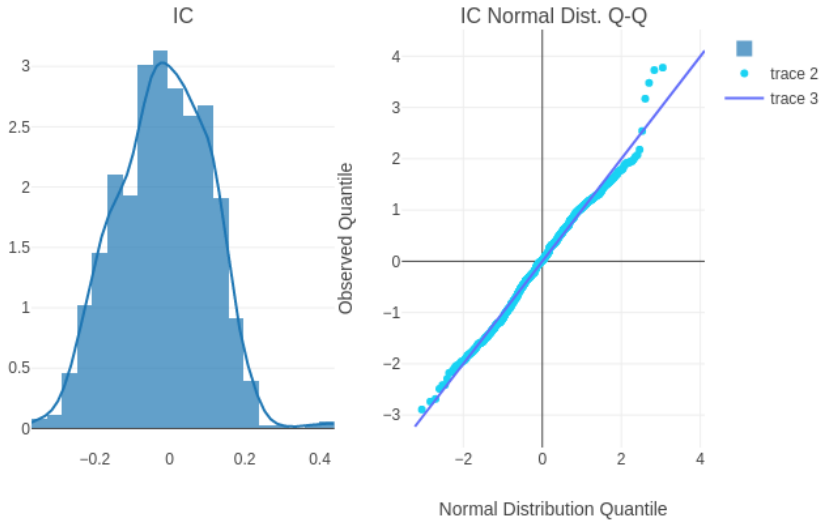


Figure 26 Distribution and normality test plot of IC

The last chart shows the autocorrelation of strategy returns (2017-2020), with values mostly between 0.7-0.9, indicating that the strategy has strong continuity and predictability. Although there are occasional declines, the overall high autocorrelation indicates that the strategy performs stably.

Overall, the six charts highlight the advantages of quantitative strategies: strong cumulative returns, near-normal IC distribution, and robust return autocorrelation. However, the recent negative trend in IC and seasonal fluctuations warn that the effectiveness of the strategy may decline. This analysis provides a solid foundation for evaluating and optimizing quantitative strategies and helps investors make decisions.

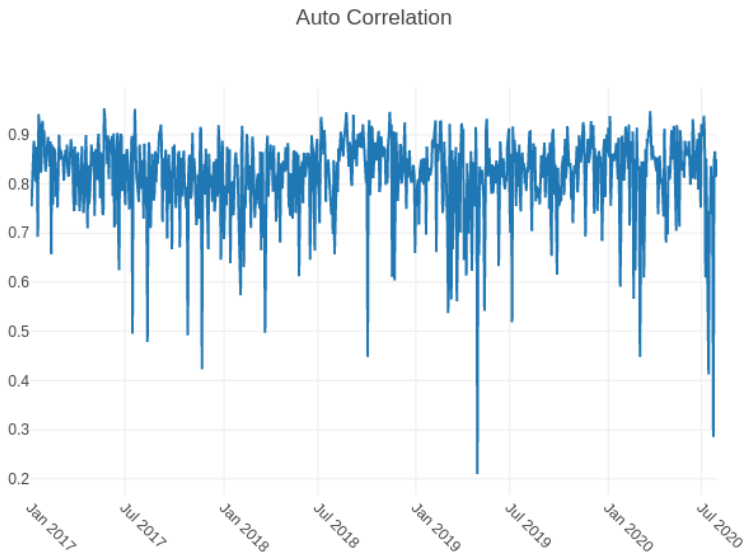


Figure 27 Diagram of Auto Correlation

References

[1] Microsoft. (2017). LightGBM: A highly efficient gradient boosting decision tree. GitHub repository. https://github.com/microsoft/LightGBM

[2] Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. ArXiv. https://doi.org/10.1145/2939672.2939785

[3] Li, T., Liu, Z., Shen, Y., Wang, X., Chen, H., & Huang, S. (2023). MASTER: Market-Guided Stock Transformer for Stock Price Forecasting. ArXiv. https://arxiv.org/abs/2312.15235

[4] Grinold, R. C., & Kahn, R. N. (1999). Active Portfolio Management: A Quantitative Approach for Producing Superior Returns and Controlling Risk (2nd ed.). McGraw-Hill.

[5] Magdon-Ismail, M., & Atiya, A. F. (2004). Maximum drawdown. Risk, 17(10), 72–75.