CMPT 733 Spring 2024 - Final Report

Integrating AI and Quantitative Analysis for Equity Investment and Portfolio Optimization

Group: HappyCNY

Shung Ho (Jonathan) Au
Hongying Yue
Sitong Zhai
Qin Duan

Motivation and Background

The intuitive explanation of Investing: When an investor, whether a seasoned investment professional or an enthusiastic amateur, commits to an equity investment, their decision is rooted in their anticipation of the company's future performance. The underlying assumption is that a solid performance by the company will be mirrored by an appreciation in its share price. The collective actions of investors' investment decisions driven by confidence in the company's prospects put upward pressure on the stock price. Over time, as these forward-looking expectations begin to unfold—or not—the speculative nature of these initial investments is rigorously evaluated against the tangible outcomes of the company's financial health, market standing, and ability to navigate the economic landscape.

However, if one assumes this process is straightforward, the truth is that a majority of actively managed funds by professionals consistently underperform the index benchmarks over the long term, revealing the complex challenges and unpredictability of predicting company success. The capital asset pricing model (CAPM) proposed by Sharpe, 1964 and Lintner, 1965 has emerged as a cornerstone in both academic and professional realms to explain asset pricing. Nevertheless, numerous research studies have documented the inefficiencies of the simple single-factor approach to the multifaceted nature of financial market dynamics.

Recent advancements and our motivations: In recent years, technological advancements have paved the way for the rise of quantitative investing. The traditional way of conducting deep dive analysis on individual stocks is time-consuming and demands substantial resources, underscoring the inevitable shift towards data-driven investment strategies. In navigating the complexities of financial markets, quantitative investing facilitates three key capabilities: 1) Analyzing larger numbers of stocks simultaneously, 2) Making decisions based on empirical evidence rather than subjective picks, and 3) Adopting a systematic framework for portfolio management that is scalable yet interpretable.

Our team combines a dynamic range of expertise, encompassing deep knowledge in finance, alongside advanced skills in statistics, machine learning, and computer science. This diverse skill set positions us uniquely to tackle the specific challenges within the domain of investment management. In line with our capabilities, our study focuses on composite stock data of the CSI 500 Index, which consists of 500 mid and small-cap stocks by market capitalization on the Shanghai and Shenzhen stock exchanges, spanning the years from 2016 to 2023.

Related work: The research conducted by Yi Fu, Shuai Cap, and Tao Pang in 2020, which utilizes back-testing to evaluate stock selection performance, shares similar methodological approaches with our study in terms of evaluation. However, our approach diverges significantly in two key areas: 1) Fixed Factor Selection Process: Unlike their methodology, which involves updating feature selection criteria across multiple periods for 50+ features, our project opts for a static evaluation of selected features. This allows us to assess the enduring value and long-term effectiveness of each factor within a consistent market context; 2) Different Stock Pool Selection and Timeframe: In contrast to their focus on the CSI 300 index, which primarily comprises large-cap stocks, our study explores the CSI 500 index, known for its mid and small-cap stocks. This selection provides a broader market view and diversifies our analysis beyond the large-cap segment. Additionally, our study extends to a more recent timeframe, concluding in late 2023. This updated period allows us to capture the latest market trends and factor performances.

Problem Statement

Questions expect to be answered:

The quantitative investing framework: In the pursuit of enabling investors to optimize stock portfolio performance without being affected by bias or emotional influences, **how** to construct an effective multifactor model that is capable of identifying top-performing stocks and making data-driven selections periodically?

The construction of representable stock features: The effectiveness of a quantitative investing model significantly depends on the quality of data and the degree to which features derived from publicly available information can accurately and meaningfully represent a company's financial health and market potential**. What** financial metrics or other trading data can be used as features for constructing multifactor models to predict stock returns? How can the selection of features and their true validity be confirmed through economic theories and interpretable artificial intelligence?

Evaluation of model performance: How can we objectively evaluate the predictive power and the performance of the model? Which benchmarks should be compared against to reveal the effectiveness and efficiency of our multifactor models in real-world conditions?

Challenges need to be addressed:

Numerous decisions to be made: The diversity of approaches and the numerous "small" decisions must be made before the model construction process. For instance, a regression model could be developed that focuses on predicting future stock prices based on historical price trends. Or, a classification model could be designed to identify outperforming stocks relative to their peers. Each choice influences the entire project trajectory, which includes initial steps such as data collection and extends to later stages like labelling for training. The highly interdependent nature of these decisions necessitates careful upfront planning, as any misstep can significantly impact the feasibility of revisiting and adjusting previous stages.

Finance and accounting knowledge gap and data quality issue: Moreover, extracting useful features requires considerable knowledge of finance and accounting. It is also very common for financial data to contain errors, missing values and extreme values. Such data quality issues may distort the model's learning process if they are not handled properly. Also, stock price data are extremely noisy and include many randomness that cannot be easily interpreted and exploited for profit (Lo & MacKinlay, 1999).

Data Science Pipeline

Stage	Key Factors Contribute to High performance
Data Collection	The model's input comprises long-term and comprehensive financial data transformed into features beyond features related to stock prices.
Data Preparation & Feature Engineering	In the feature preprocessing stage, we conducted outlier removal and industry-neutralization processing. Features were selected based on their significance.
Data Transformation & Labelling	During labeling, we selected the top 30% performers and bottom 30% performers, excluding the middle portion. Therefore, the model is more sensitive to changes in features.
Model Training	Multiple models are trained. XGboost is selected as one of the best performing model.
Model Evaluation	The evaluation scenario is designed to adjust the portfolio monthly and calculate long-term returns.

1. **Research and Planning:** We established our goal to develop a stock classification model capable of predicting whether a given stock will achieve excess returns compared to the index benchmark. The first step involves understanding how public accounting data is traditionally utilized within the industry to assess a company's performance. Through this exploration, we found that financial health and stock momentum are heavily considered by financial analysts, who may place different emphases based on their "style" or preferences. To align better with our motivation to eliminate personal bias and simultaneously evaluate numerous stocks that would traditionally require manual assessment, our primary mission became the selection of fundamental factors that provide a comprehensive view of a company's condition. We opted for a strategy that relies on explicit financial features, rather than indirect methods like sentiment analysis from news reports etc. This direct approach ensures our model is rooted in the financial fundamentals driving stock performance and allows us to derive insights from an economic standpoint. We identified 5 categories of indicators that can potentially be constructed as features in the modelling phase. We then evaluated various data sources to ensure the availability, reliability, and relevance of the data needed to accomplish our goals.

2. **Data Collection**:

    2.1. **Scope and Timeframe:** This research utilized the JoinQuant API, an open-source comprehensive financial data platform that offers access to a wide array of stock financial and trading information. We queried the composite stocks under the CSI 500 Index with a timespan from January 1, 2016, until December 31, 2023, to ensure the timeframe is long enough to capture multiple market and economic cycles. The CSI 500 composite stocks comprise 500 small to medium-market capitalization stocks traded on the Shanghai Stock

Exchange and the Shenzhen Stock Exchange. The list of composite stocks is refreshed semi-annually based on their most recent market capitalization, which is the total value of all a company's shares of stock.

2.2. **Data Overview:** The data collected at this stage is raw and extensive. Overall it comprises the following elements.

- 2.2.1. Trade Day: *Dates* - only query data for market operating days between 2016-2023.
- 2.2.2. Stock Identifiers: *Strings* - identifying each stock uniquely.
- 2.2.3. Financial Metrics (multiple): *Floats* - This segment includes raw monetary values sourced from balance sheets and income statements, such as *$ short-term liabilities* and *$ cash equivalents*, which were subsequently transformed into ratios for evaluating financial health. The rationale behind the selection and utilization of each metric will be elaborated further in the Methodologies section.
- 2.2.4. Technical analysis Indicators (multiple): *Floats* - Incorporates calculated indicators on a rolling basis, such as RSI, MACD, which provide insights into stock momentum based on price movements.
- 2.2.5. Price Data: *Floats* - Contains the stock's closing price for a trading day.

2.3. **Initial Filtering:** To refine the integrity and quality of the dataset, we imposed the following filtering conditions when querying the data, because in practice such a list of stocks would not be a favourable option to consider:

- 2.3.1. Remove stocks marked as "ST" that indicate special treatment due to severe financial distress.
- 2.3.2. Remove stocks that paused trading on a given trading date.
- 2.3.3. Remove newly traded stocks (under 90 days) that lack a sufficient historical record.

3. Data Preparation and Feature Engineering:

3.1. **Limit Extreme Values:** To reduce the impact of outliers in the features that could skew the model training, we employed a statistical method known as winsorization that capped the feature values at 3 standard deviations from the mean, thus limiting both the upper and lower extremes.

3.2. **Interpolate Missing Value with Industry Mean:** Roughly 3% of the data contained null values in some financial metrics data. We imputed these gaps with industry means to preserve the industry-specific characteristics within the data.

3.3. **Feature Bias Reduction and Normalization:** To ensure a more equitable comparison across all stocks, we took into account the industry and scale influence of different stocks. The objective was to isolate the "intrinsic" value of the factors across various stocks by performing an Ordinary Least Squares regression. The residual represents the portion of the factor's variation not explained by size or sector, thus "neutralizing" these effects to eliminate biases introduced by company size and industry differences. For example, banking stocks typically exhibit lower price-to-earnings (P/E) ratios compared

to technology stocks due to differences in business models and growth expectations. A low P/E ratio in banking might not indicate the same value opportunity as it would in tech because the sectors inherently operate with different financial structures and market dynamics. By applying OLS regression and focusing on the residuals, we neutralize these sector and scale effects.

3.4. **Feature Standardization:** Features containing various financial and trading metrics involve different scales and ranges widely. For instance, earnings can range from billions while sales growth is in percentages, or stock price volatility is not in any way comparable to price-to-earnings ratios etc. This disparity in scale might unduly weigh larger values or overemphasize the significance of smaller ones. We adopted the z-score standardization technique, resulting in a distribution of the data with a mean of 0 and a standard deviation of 1.

4. **Data Transformation**:

4.1. **Monthly Return Calculation:** Each stock's return is measured by the percentage change in closing price at the end of the month compared to the start of the month.

4.2. **Data Labelling:** Once the monthly returns are calculated, the data is sorted in descending order based on these returns. This ordering allows the identification of stocks that performed the best and the worst during the month. We labelled the top 30% of stocks as having positive performance (1) and the bottom 30% as negative performance (0). The rationale behind this approach is to train a model that predicts which stocks are likely to be top performers.

5. Model Training:

5.1. Train-Test dataset split: we allocated trading data from January 1, 2016, to December 31, 2019, for model training purposes. Subsequently, data from January 1, 2020, to December 31, 2023, was designated for testing.

5.2. Feature selection: We have a total of 20 features. To select the most useful features for training, in our best version of the model, we applied the embedded method using random forests to compute the feature importance of each feature. We set the threshold to be 0.05, which means we only used the features that have at least 0.05 for the model training.

6. **Model Evaluation and Backtesting:** The model will be evaluated based on its predictive performance and its portfolio return. A hypothetical portfolio is constructed by "buying" stocks classified as $y = 1$ based on the model's monthly predictions. This simulates a proactive investment strategy that dynamically adjusts the composition of the portfolio at the **beginning** of each month. The simulation repeats this process each month during the test period, thereby effectively rolling the model's forecasts and adjusting the portfolio accordingly. This rolling forecast methodology reflects a realistic investment scenario where decisions are made based on the latest available data, at least on a monthly basis.

Methodology

1. **Features Summary**: In our model, we aim to represent a stock from as comprehensive a perspective as possible, using five main categories of features. Each category is designed to capture a different aspect of a stock's performance and market position, providing a holistic view.

    a. **Valuation Metrics**: These metrics assess how expensive a stock is relative to various financial fundamentals, reflecting market expectations for a company's growth and profitability.
        i. PE ratio: Stock price to Earnings. Reflects how much investors are willing to pay per dollar of earnings
        ii. PB ratio: Stock price to Net Asset
        iii. PS ratio: Stock price to Sales Revenue
        iv. PCF ratio: Stock price to Cash Flow

    b. **Financial Leverage:** Indicates the extent to which a company uses borrowing to finance its operations, with higher leverage pointing to greater use of debt.
        i. Debt to Equity Ratio: Measures liability over total assets
        ii. Cash ratio: Measures short-term liabilities over cash at hand
        iii. Current ratio: Measures short-term liabilities over liquid assets

    c. **Profitability Metrics**: Higher profitability suggests a company is efficient in converting sales into actual profits.
        i. Gross Profit Margin: (Sales Revenue - raw material cost) / Sales Revenue
        ii. Net Profit Margin: (Sales Revenue - raw material cost - operating cost) / Sales Revenue
        iii. Adjusted Profit to Profit: Measures the proportion of profit from the company's primary business, low p-to-p means the company is not focusing on its main business. (eg: a car manufacturer has investment income from real estate)

    d. **Growth:** Measures the percentage growth in profit over time
        i. Inc_total_revenue: Measures % increase in revenue
        ii. Inc_net_profit: Measures % increase in net profit
        iii. Inc_operating_profit: Measures % increase in operating profit

    e. **Momentum:** Captures trading activities that measure short-term stock price movement relative to average to identify momentum shifts.
        i. HSL (Turnover rate): Indicates the frequency of a stock bought and sold over the past week. High HSL means an extremely short holding period for a unit of stock.
        ii. DEA: Moving average difference between short-term price trend (10 days) versus long-term price trend (30 days), over 15 days.
        iii. BIAS: Compares the current stock price to its average over 20-day average to identify deviations from the typical level.

2. **Experiments and Evaluation Techniques:**
    a. **Backtesting and benchmark comparison:** We apply the model's decision-making criteria to past market data (2019-2023) to see how it would have performed, providing insights into its predictive accuracy and robustness. We compare key performance metrics, including *'Cumulative Return', 'Annualized Return', 'Maximum Drawdown', 'Proportion of Months Beating the Benchmark', and 'Proportion of Positive Return Months'* against the index benchmark (CSI 500). This comprehensive comparison enables us to evaluate not just the raw performance but also the risk-adjusted returns, resilience during market downturns, and the model's consistency in outperforming the market benchmark.
    b. **Rank-Based Backtesting:** Recall that the classification model aims to predict which stocks are likely to be top performers. In backtesting, after predicting which stocks are likely to perform well, we ranked the stocks according to their predicted likelihood of and then segmented them into quintiles, from highest to lowest confidence. By simulating and comparing the returns across these quintiles/portfolios, we evaluate the model's proficiency in distinguishing between high and low performers. Ideally, an effective model should demonstrate that stocks in the top quintiles consistently outperform those in the lower quintiles, validating the model's predictive accuracy and its utility in guiding investment decisions.
    c. **Time-Series Cross-Validation:** The F1 score provides insight into the balance between recall and precision. We utilized the F1 score to assess our model's performance consistency over time, tracking how well it maintains accuracy in predicting stock performance across different market conditions and identifying any potential degradation in performance in changing market dynamics.

3. **Choice of Training Algorithms for experiments:**
    a. **Gaussian Naive Bayes:** The Naive Bayes Classifier is known for its simplicity and speed in training, making it an ideal starting point for initial model testing. It selects the class with the highest posterior probability given the features, which is straightforward and grounded in probability theory. We use this classifier for initial testing due to its minimal tuning requirements, offering us a relatively simple yet effective probabilistic estimation of our model's performance. This initial step helps us gauge the viability of our features and determine whether our feature engineering is steering us in the right direction before committing to more complex models.
    b. **XGBoost (Extreme Gradient Boosting):** We proceeded to experiment with XGBoost to explore the potential benefits of a more sophisticated model. Employing a more complex model like XGBoost helps ensure that our investment modelling strategy is more robust, comprehensive, and capable of adapting to the nuances of financial data. We applied 5-fold cross-validation and observed that a “shallower” tree structure, moderate learning rate, and small size ensembles enhance its generalization capabilities. The hyperparameter used in the following analysis is: *{'learning_rate': 0.2, 'max_depth': 1, 'n_estimators': 10}*.

4. **Bias Reduction**: Recall our data preprocessing step involves multiple operations, including Limiting extreme values, Neutralization (reducing feature bias across various stocks due to differences in company size and industry), and standardization (eliminating disparity in scale due to various feature sizes). By comparing two SHAP plots—one with raw data and the other with preprocessed data—we discerned substantial differences in how the XGBoost model interpreted feature importance(Figures 1 and 2). Prior to preprocessing, the model was susceptible to being skewed by outliers and biased by disparities in feature scales, as was evident with features like HSL and PS Ratio. Post-preprocessing, the impact of these features became more consistent and uniform across the model's predictions. This suggests that the preprocessing steps helped the model to focus on the true signals within the features, rather than being misled by noise or biases. These findings affirm that our preprocessing efforts were crucial in distilling the data to its truest form.

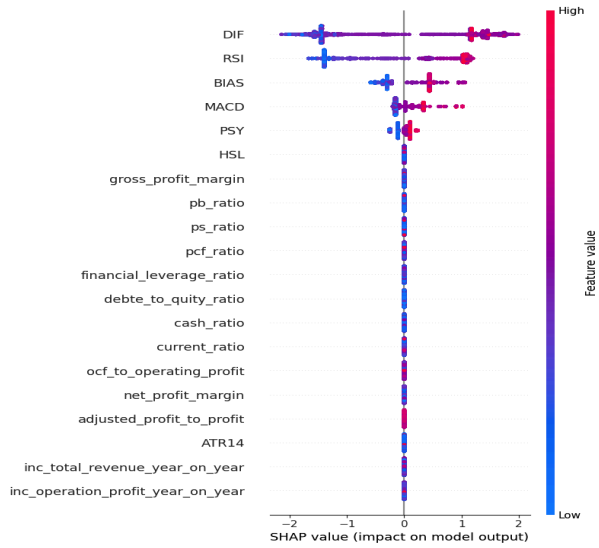


Figure 1: SHAP-raw data

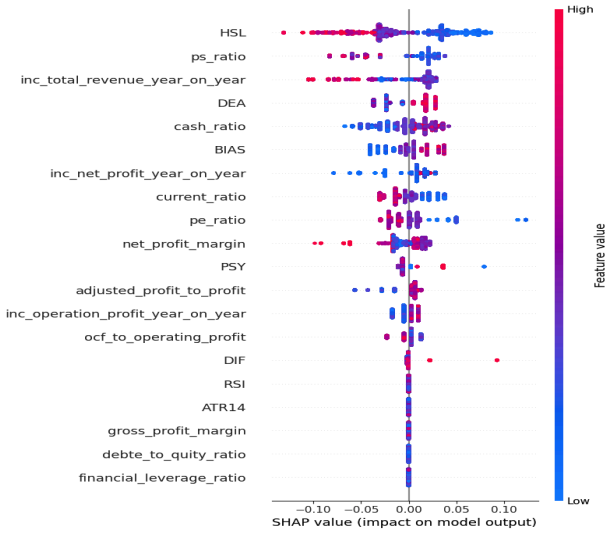


Figure 2: SHAP-preprocessed

Based on the model, we observed the following properties that have a higher chance of being good performers:

1. **Low valuation metrics (eg: low ps_ratio, pe_ratio)**
   Low PS and PE ratios may signal stocks poised for growth yet to be realized by the market. These ratios suggest that the stock could be undervalued, offering a potentially attractive entry point for investors seeking value.

2. **Moderate growth in total revenue and net profit (extremely high growth impacts negatively)**
   Steady, moderate growth is often perceived as sustainable, marking a company as a stable investment in the volatile Chinese market. Excessive growth rates, while alluring, can be unsustainable and risky over the long term.

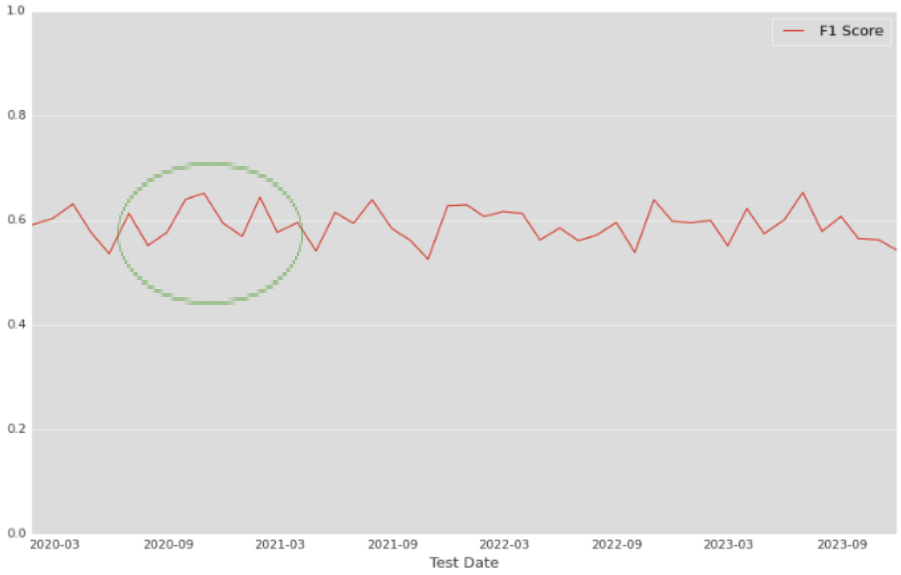
3. **Long average holding period of a stock over the past week. (eg: low HSL means the average turnover rate of a unit of stock is not extremely frequent)**
   A low HSL indicates investor confidence and commitment to a stock's long-term value in China's retail-driven market. It implies stability and a reduction in speculative trading, which is favoured by value investors

4. **Upward momentum shift in price trend (eg: High DEA, BIAS)**
   Upward trends in DEA and BIAS suggest increasing market recognition and a potential rally in stock prices. This positive momentum is attractive in China's market, where such shifts can amplify through high investor engagement.

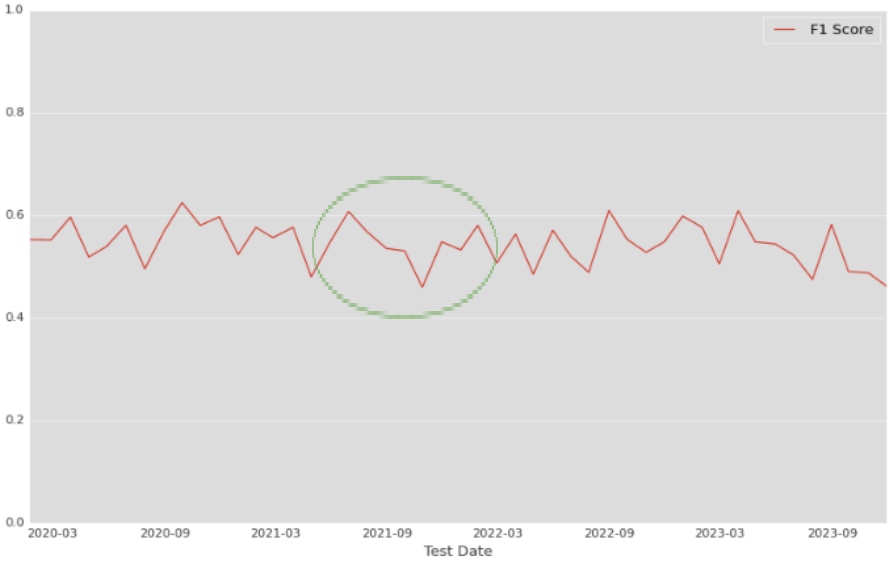
Evaluation

1. **F Score over time:**

   a. Naive Bayes:



b. XGBoost *(learning_rate': 0.2, 'max_depth': 1, 'n_estimators': 10)*



Insight 1: Both the simple Naive Bayes and the more sophisticated XGBoost classifiers demonstrated consistent performance over time.

In analyzing the **F1 scores** from January 2019 to December 2023, the Naive Bayes model averaged 0.59 and XGBoost 0.54. Despite moderate fluctuations between 0.5 to 0.6, there was no evident declining trend in performance over the test period, meaning that the models could maintain a a certain level of consistency in their predictions. A notable dip in the F1 score occurred in 2021-09 in both models, coinciding with a significant downward shift in the market reflected in the index benchmark (CSI 500), meaning that the performance would be affected by the sharp market shift. We will proceed to examine the return performance metrics to assess the real-world implications of these findings.

2. Benchmark comparison:

a. Naive Bayes:

Metrics	Top 50%	Bottom 50%	Benchmark (CSI 500)	definition
Cumulative Return	42.4%	25.7%	1.0%	Jan 2020 - Dec 2023 return %
Annualized Return	9.44%	6.01%	0.25%	Annualized return %
Maximum Drawdown	15.18%	20.99%	26.23%	Maximum fall in value %
Proportion of Months Beating the Benchmark	57.45%	59.57%	NA	Proportions of months have higher return than CSI 500
Proportion of Positive Return Months	57.45%	57.45	51.06%	Proportions of months with +ve return
Information Ratio	4.19	2.89	NA	Excess return / Standard deviation

b. XGBoost:

Metrics	Top 50%	Bottom 50%	Benchmark (CSI 500)	definition
Cumulative Return	**43.2%**	25.0%	1.0%	Jan 2020 - Dec 2023 return %
Annualized Return	9.6%	5.86%	0.25%	Annualized return %
Maximum Drawdown	15.99%	20.22%	26.23%	Maximum fall in value %
Proportion of Months Beating the Benchmark	**61.7%**	55.32%	NA	Proportions of months have higher return than CSI 500
Proportion of Positive Return Months	**59.57%**	55.32%	51.06%	Proportions of months with +ve return
Information Ratio	**4.8**	2.62	NA	Excess return / Standard deviation

Insight 2: Both models were able to outperform the CSI 500 Index despite the Chinese A-Share market facing significant challenges after 2022. Demonstrating superior returns and exhibited reduced risk

Model's cumulative Return vs Benchmark (CSI 500):
After predicting which stocks are likely to perform well, those stocks are then segmented into two quantiles based on the model's confidence in their performance, with the Top 50% containing the stocks with the highest confidence of good performance and Bottom 50% containing those with lower confidence **among the predicted good performers.** Both models performed similarly in returns and have the ability to limit maximum drawdown. The top 50% quantile of both models outperformed the benchmark (CSI 500) return by roughly 40%, and the maximum drawdown was 11% less compared to the benchmark (CSI 500). Maximum drawdown is a key metric that indicates the risk of a portfolio investment by measuring the maximum fall in % value over a specific period.

Top 50% quantile vs Bottom 50% within each Model:
We can also observe that the Top 50% quantile has better return and lower maximum drawdown compared to the bottom 50% stock quantile in both models, meaning that the models were **effective in distinguishing between stocks with higher and lower potential performance.**

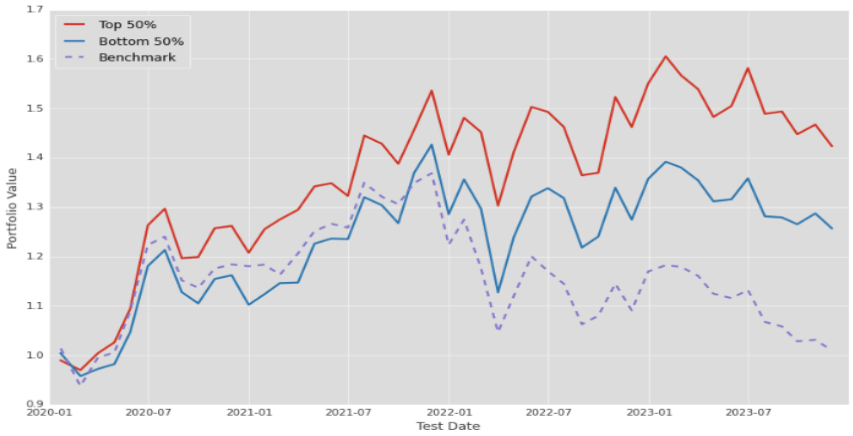
Proportions of months beating benchmark:
Additionally, we looked at the proportion of months in which the model yields higher monthly returns than the benchmark because it tells the odds of outperformance from a probabilistic perspective. We found that the **XGBoost model surpassed the index benchmark in 61.7% of the months** throughout the testing period, with 59% of the months showing positive returns, compared to 51.1% for the index benchmark. In this regard, XGBoost slightly outperformed the Naive Bayes model.

Excess return relative to standard deviation:
Information ratio is a common metric to evaluate the risk-adjusted return of a portfolio compared to the baseline (CSI 500). We observed the highest Information Ratio in the XGBoost model (Top 50%), amounting to 4.8. The Information Ratio is a metric that assesses **excess returns while considering the standard deviation**; a higher ratio signifies superior excess returns relative to the risk incurred, denoted

by a lower standard deviation. The Naive Bayes model also performed commendably, with the Top 50% achieving a ratio of 4.19.

4. Rank-based backtest - top 50% vs bottom 50%

1. Naive Bayes:



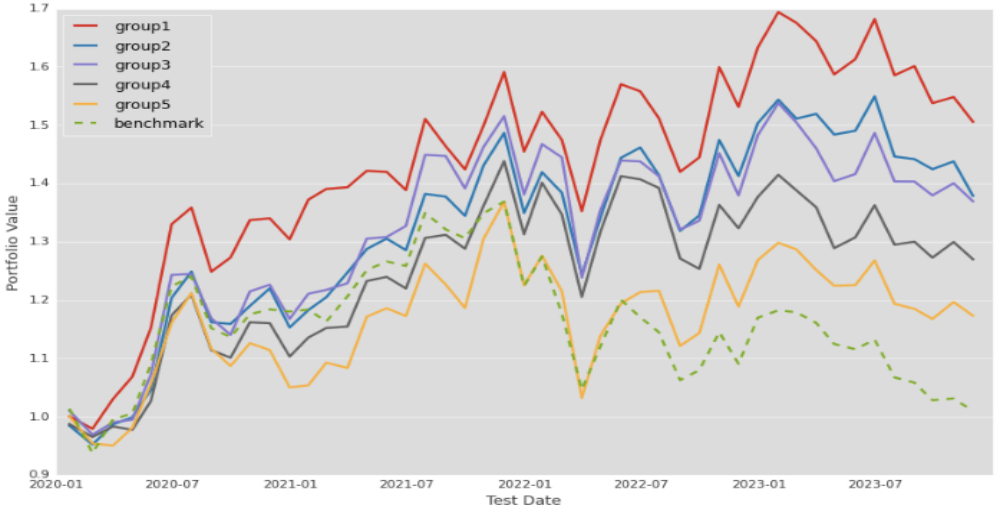
2. XGBoost:

Insight 3: Both models exhibited remarkably consistent patterns, with the top 50% quantile consistently outperforming the bottom quantile. This indicates that stocks predicted to be good performers indeed exhibited superior performance compared to those with lower confidence, further affirming the models' effectiveness. Notably, the top 50% model demonstrated enhanced resilience during market downturns in 2022-01.

5. XGBoost Deep Dive (5 groups)

Metrics	g1	g2	g3	g4	g5	Benchmark (CSI 500)	definition
Cumulative Return	**50.6%**	37.9%	36.9%	27.0%	17.3%	1.0%	Jan 2020 - Dec 2023 return %
Annualized Return	**11.02%**	8.55%	8.35%	6.29%	4.16%	0.25%	Annualized return %
Annualized Excess Return	**10.55%**	8.02%	8.1%	5.74%	3.87%	NA	Annualized return % compared to CSI 500
Maximum Drawdown	**15%**	16.39%	18.29%	16.2%	24.58%	26.23%	Maximum fall in value %
Proportion of Months Beating the Benchmark	57.45%	**70.21%**	61.7%	59.57%	55.32%	NA	Proportions of months have higher return than CSI 500
Proportion of Positive Return Months	57.45%	57.45%	57.45%	51.06%	53.19%	51.06%	Proportions of months with +ve return
Information Ratio	5.04	4.0	3.71	2.56	1.46	NA	Excess return / Standard deviation

We further segmented the XGBoost portfolio into five quantiles to evaluate whether our insight about the model's confidence in effectively isolating good performers is consistent with previous results. Our findings are that **g1 (top 20%) has the best performance overall**, and the higher quantiles always performed better than the lower quantiles in terms of cumulative return and maximum drawdown.



Data Product

Our data product represents a refined stock classification system meticulously crafted to discern superior stocks from the CSI 500 Index, fostering the construction of high-performing portfolios that outshine the index and predict excess returns. Rooted in robust financial fundamentals, it harnesses explicit financial metrics extracted from publicly available accounting data to gauge company performance.

Powered by the JoinQuant API, our system facilitates comprehensive data gathering, while an exhaustive preprocessing script meticulously handles outliers, missing data values, neutralization, and standardization, ensuring the integrity and uniformity of the analysis prior to training. At its core lies the XGBoost algorithm, empowering portfolio managers to sift through a composite of 500 stocks and pinpoint those poised for optimal performance.

This data product boasts the agility to promptly adapt portfolios based on algorithmic insights, facilitating responsive updates guided by market dynamics. By dynamically adjusting the portfolio monthly in accordance with predicted favorites, our product delivers investment flexibility alongside impartial and resilient evaluations.

In essence, our solution empowers investors to seize market opportunities with confidence grounded in evidence, mitigating emotional biases and fostering a data-driven approach to investment decision-making.

Lessons Learned

The importance of data preprocessing truly stands out in the learning process, especially when dealing with financial information. It's not just about feeding data into an algorithm and hoping for the best. To really get it right, we need to dive deep into the specific quirks and biases of the financial market. This involves adjusting our data carefully—like ensuring all the numbers are comparable across different companies—before we start analyzing it. If we skip this step, the model could end up making decisions based on skewed information. So, preprocessing is not just about cleaning up data; it’s about making sure our analysis is fair and accurate by removing any potential biases right from the start.

We discovered a profound paradox in developing and evaluating both simple and complex models: despite the allure of complex algorithms, the performance difference between simple models like Naive Bayes and complex models like XGBoost is not as significant as one might think. This observation is a powerful reminder to follow Occam's Razor when choosing a model - complexity does not always equate to superior performance. On the contrary, the clarity and interpretability of a simple model can often provide equally valuable insights, especially given the vagaries of financial markets. As we blend the analytical rigor of data science with a nuanced understanding of financial dynamics, this project has furthered our understanding that striking the right balance between simplicity and complexity, underpinned by robust data preprocessing, is critical to the development of effective data-driven investment strategies.

Summary

Our initiative aimed to enable the simultaneous analysis of numerous stocks to support data-driven investment decisions and build a stock portfolio to outperform index. Faced with the challenges of market unpredictability and the considerable effort required to analyze and curate features for enhanced

predictive power, we devoted significant effort to ensuring the acquisition of the right quantity and quality of features during the project's initial phase. We also delved into feature importance and SHAP plots to comprehend how our model bases its decisions on each feature, achieving a logical understanding from an economic perspective.

We recognize that we allocated more effort towards curating and selecting features than to deploying or fine-tuning more complex algorithms, which proved to be rewarding. Leveraging our expertise in finance and statistics, we preprocessed the data to eliminate potential biases that could significantly affect learning, thus allowing the model to concentrate on signals from the features.

To mitigate overfitting and reduce the risk of project failure, we chose not to predict specific price points for individual good performance stock or below-average stock based on their relative performance to peers using monthly data. Initially in training, we utilized classical, explainable, and less complex models such as Naive Bayes. Subsequently, we employed XGBoost to assess the marginal gains in performance, observing a slight improvement. However, even the optimal hyperparameters for the XGBoost model indicated a preference for a very shallow tree structure, to prevent severe overfitting and underperformance compared to benchmarks. This implies that, given the current quality and quantity of features, the bottleneck lies not within the algorithm. Future work should aim to expand the feature scope before deploying more complex algorithms capable of discerning more complex representations.

For evaluation, we conducted portfolio simulations, constructing hypothetical portfolios based on the model’s predictions and change the portfolio monthly based on latest prediction. Our portfolio successfully outperformed the index benchmark in terms of percentage return and maximum loss over the test period from 2020 to 2023. We validated the model's effectiveness through backtesting across multiple probability quantiles, showing that portfolios comprising stocks with the highest probability of good performance consistently outperformed those with lower probabilities. We divided the probability quintiles into five detailed intervals, and the results matched the expected performance. This confirms that the model's predictions are not random; they can indeed distinguish between good, average, and below-average stocks, as we observed distinct differences in return over an extended period.

Interestingly, we discovered that using one less year for training data (2016-2018 for training and 2019-2023) for testing resulted in performance that closely matched the benchmark. The model did not outperform in bullish market conditions but showed greater resilience during downturns, indicating lower risk in declines and less aggressiveness in rising markets. However, this insight was secondary; once we incorporated an additional year of training data, the model was able to outperform in both rising and falling markets, which then became our primary focus.

As potential next steps, extending the time frame further, for example from 2005 to 2024, could help determine if insights remain consistent with additional data. Applying the same methodology to other stock composites, such as the CSI 300 or S&P 500, could offer insights into different market characteristics and predictive performance. Retraining the model to ascertain if performance can be sustained with a similar methodology also presents a viable option.

END