Integrating AI and Quantitative Analysis for Equity Investment and Portfolio Optimization

SFU

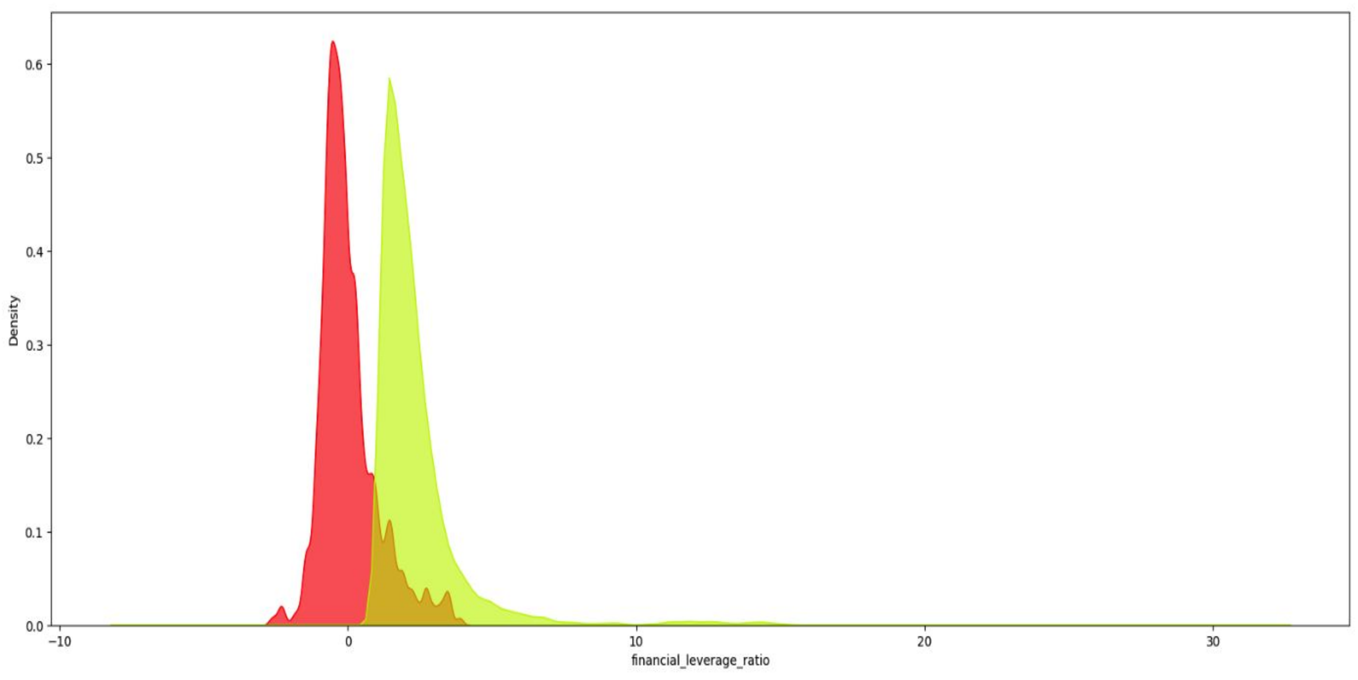
Jonathan Au, Hongying Yue, Sitong Zhai, Qin Duan

Motivations

1. Develop a scalable yet effective machine learning framework that can identify high-return stocks to guide investment decisions
2. Eliminate emotional bias and adopt effective features and algorithm to construct investment portfolio
3. Outperform the index benchmark in terms of **%** return and risk (maximum loss %)

Data Preprocessing Pipeline

1. **Limit Extreme Values** by capping feature values at 3 std from the mean
2. **Interpolate Missing Values** with Industry Mean
3. **Feature Bias Reduction** using OLS regression to extract residuals and eliminate bias introduced by company size and industry differences.
4. **Z-Score Standardization** to eliminate disparity in feature scale size



Methodologies

Data Collection → Data Preparation & Feature Engineering → Data Transformation & Labelling → Model Training → Model Evaluation

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.

Data: API Query from Joint Quant Data, covering Jan 2016 to Dec 2023
Stock Universe: CSI 500 China's Mid & Small-cap Universe
Binary Classification model: The classifier is trained to predict whether a given stock is likely to be top performers. Then hypothetical portfolios can be constructed by "buying" stocks classified as y=1 based on the model's monthly predictions, or based on some confidence thresholds. The process repeats at the beginning of each month during the test period, thereby rolling the model's forecasts and reflect an investment scenario when decisions are made based on the latest available data on a regular basis (assumed monthly).
Train: <u>2016-01 to 2019-12</u> **Test:** <u>2020-01 to 2023-12</u>
Features Summary:

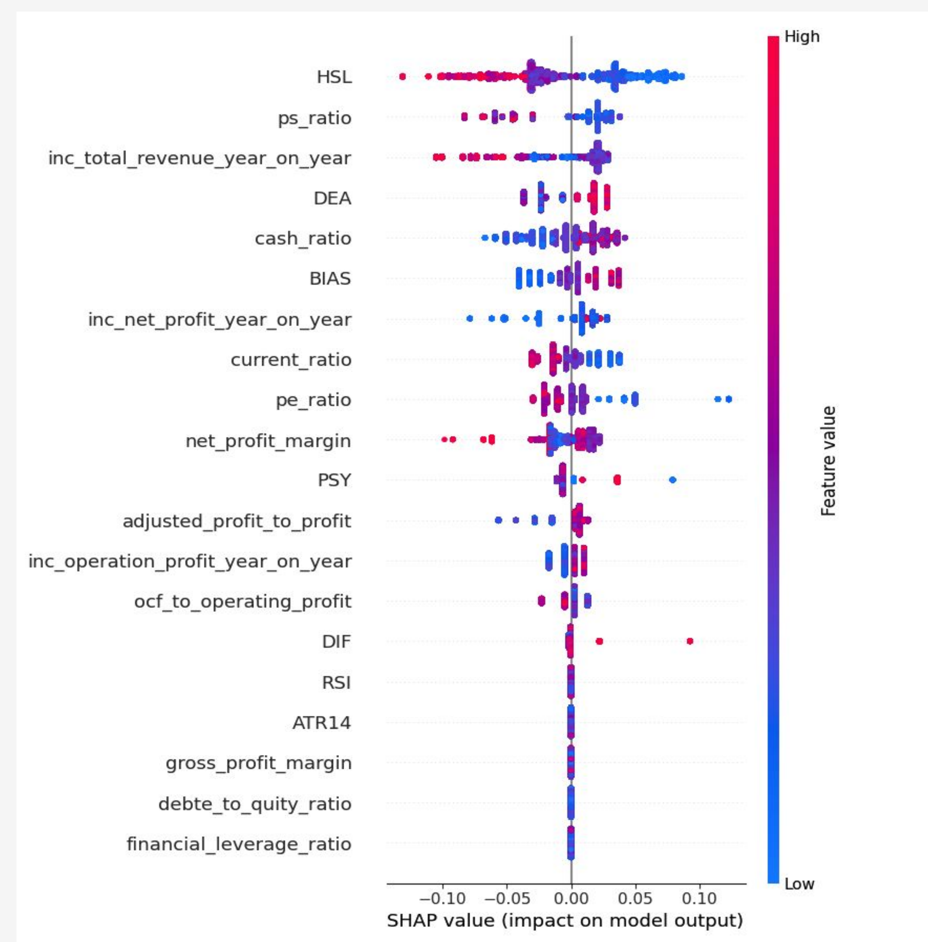
- 4 Valuation Metrics: *pe ratio, pb ratio, ps ratio, pcf ratio* to evaluate expensiveness of price
- 3 Financial Leverage Metrics: *debt to equity, cash ratio, current ratio* to evaluate debt level
- 3 Profitability Metrics: *Gross profit margin, Net profit margin, Adjusted profit to profit*
- 3 Growth Metrics: *% increase in total revenue, net profit , operating profit*
- 3 Momentum Indicators: *HSL, DEA, BIAS* showing short term/long term price movement

Evaluation and Results

Rank-Based Backtesting and Benchmark Comparison:

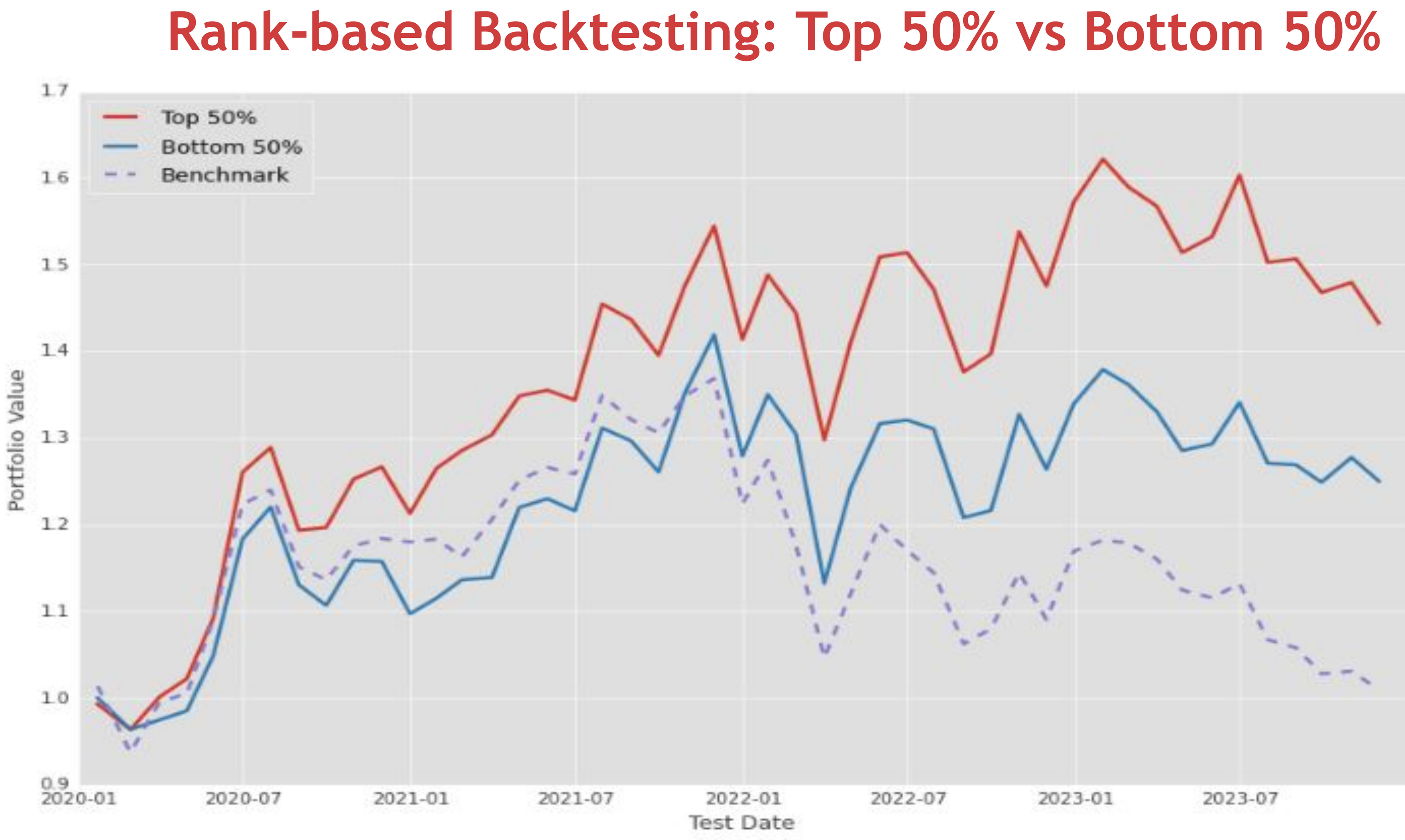
- Segmented the **XGBoost** portfolio into five quantiles based on the **model's predicted probabilities of good performance** eg: *g1= top10% … g5 = 40-50% quantile* in probability
- The top quantile of the models outperformed the benchmark (CSI 500) return by roughly **49%**, and the maximum drawdown was **11%** less compared to the benchmark.
- XGBoost model surpassed the index benchmark in **57.5% of the months** throughout the testing period 2020-2023.
- Upper quantiles always have better performance compared to the Lower quantiles (g1>…>g5), the models were **effective in distinguishing between stocks with higher and lower potential Performance.**

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

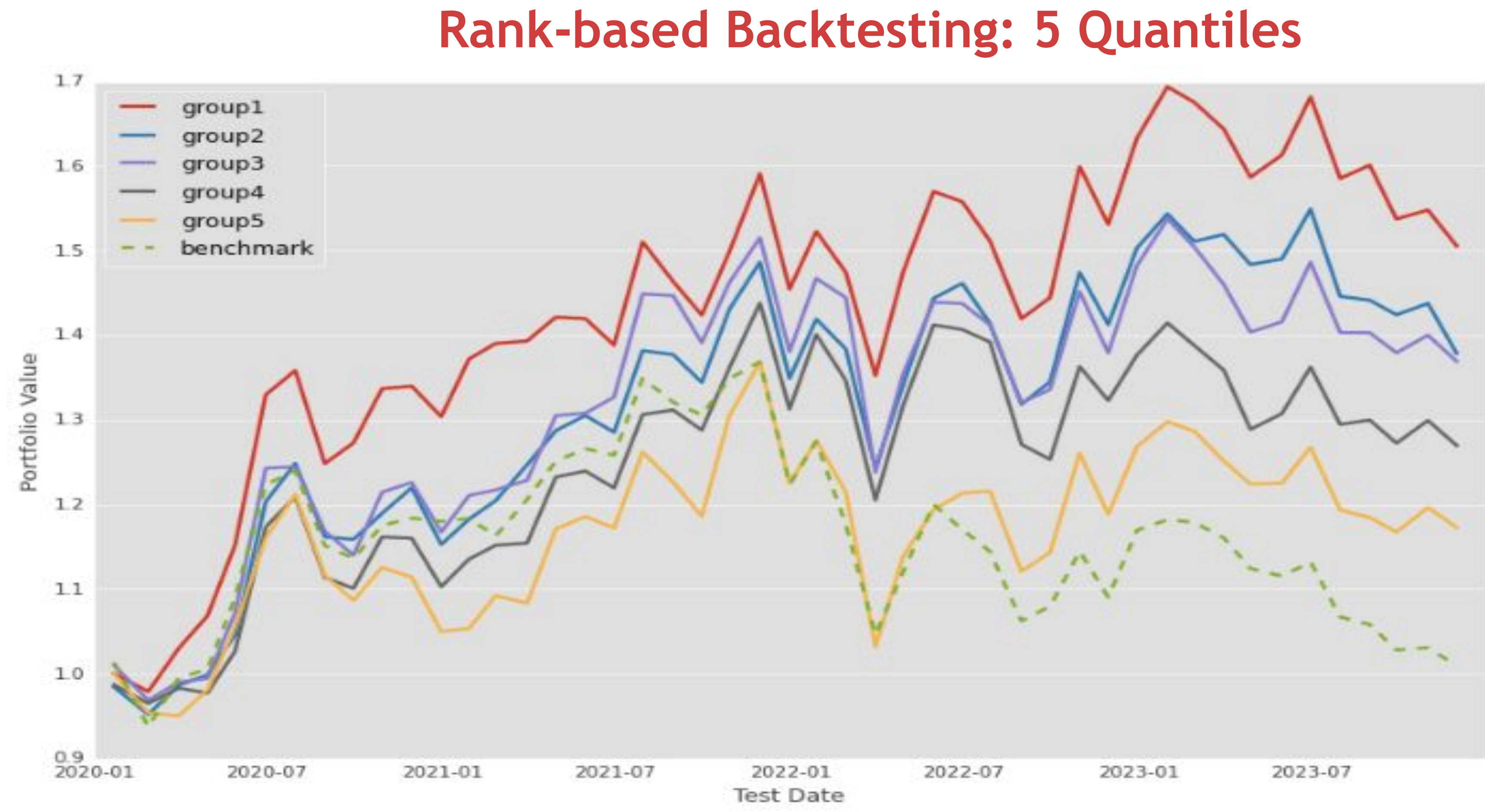


- XGBoost Model has high information ratio, meaning that the **excess returns accounted for standard deviation** is superior than benchmark index (CSI 500).

Rank-based Backtesting: Top 50% vs Bottom 50%



Rank-based Backtesting: 5 Quantiles



F-Score over time

- No evident declining trend in performance over the test period, meaning that the models could maintain a certain level of consistency in their predictions
- Average F1-Score was 0.54

Conclusion

Properties of good performers according to the SHAP plot: **low valuation** multiples, **moderate growth in total revenue** and net profit, **avoid extremely short average holding period** (HSL) of a stock, u**pward momentum shift in price** trend etc.

Stocks predicted to be **good performers indeed exhibited superior performance compared to those with lower confidence**, further affirming the models' effectiveness. The optimal confidence threshold for XGBoost portfolio is observed to be the top 10% quantile.

Simple models can also achieve good results. The same evaluation procedure applied to a **Naive Bayes Classifier** yielded results remarkably close to those of XGBoost, underperforming by a very slight margin. Additionally, the XGBoost model used hyperparameters that favored a **shallower tree structure** with a **moderate learning rate** and a **smaller ensemble size**, which provided the best generalization capabilities. Detailed findings are included in the formal report.