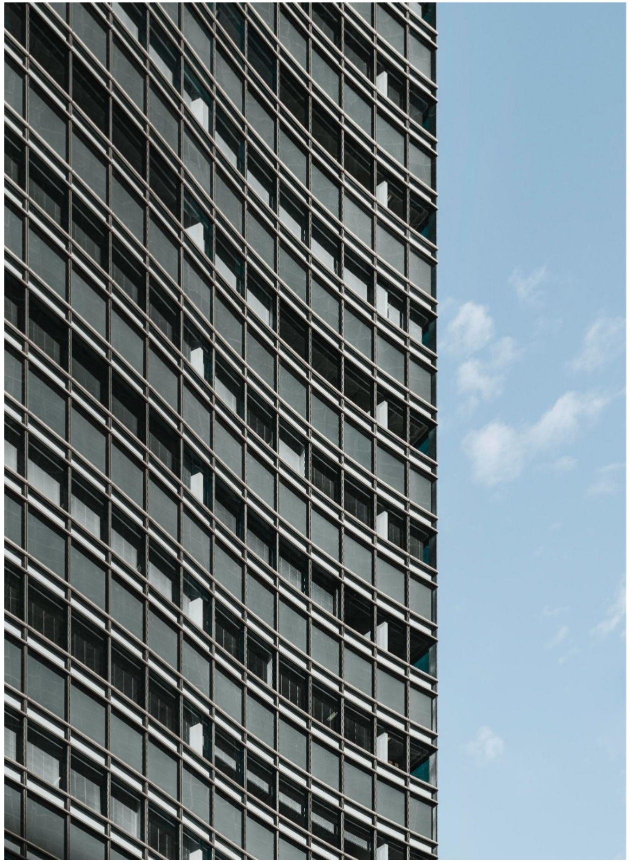
CMPT 733 Spring 2024

Integrating AI and Quantitative Analysis for Equity Investment and Portfolio Optimization

Group: HappyCNY

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



AGENDA

01 Motivation

02 Data Science Workflow

03 Methodology

04 Evaluation

Key Motivations

Current State:

1. Deep dive analysis of individual stocks is time consuming and resource intensive
2. Price movement affected by many forces/factors
3. The Majority falls behind indexes in long run

Our goals:

1. **Develop a rational and consistent Investing Framework**
   a. **Identify high-return stocks**
   b. Eliminate emotional Bias
   c. Scalable

1. **Enhance Return Potential through Machine Learning**
   a. Factor Investing
   b. **Outperform index benchmark in risk-adjusted return**
   c. Model interpretability

Data Science Workflow

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

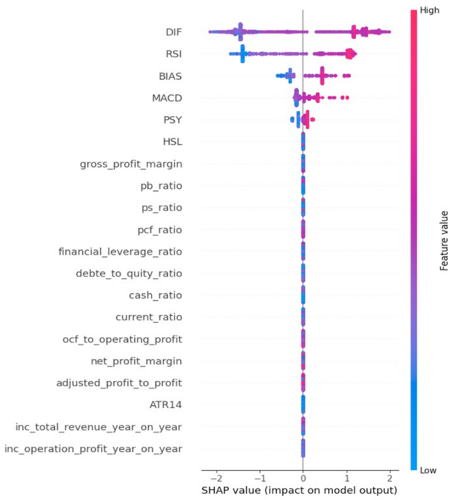
Stage	Key Factors Contribute to High performance
Data Collection	The model's input comprises long-term and comprehensive financial & trading data transformed into features beyond features related to stock prices.
Data Preparation & Feature Engineering	In the feature preprocessing stage, we conducted outlier removal and industry-neutralization processing. Features were selected based on their significance.
Data Transformation & Labelling	During labeling, we selected the monthly 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 JointQuant Data
Stock Universe: CSI 500 - China's Small-Mid capitalization A-Shares
Train: 2016-01 to 2019-12
Test: 2020-01 to 2023-12
Features: 5 big categories ***valuation***, ***leverage***, ***profitability***, ***growth***, ***momentum*** with a total of 16 features
Model Type: Binary Classification
Rolling-forward Backtesting: Buying stocks at the beginning of each month based on monthly prediction between 2020 - 2023

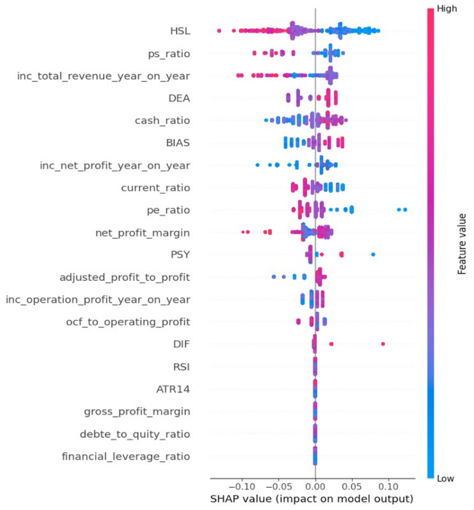
Methodologies

Data Preprocessing:

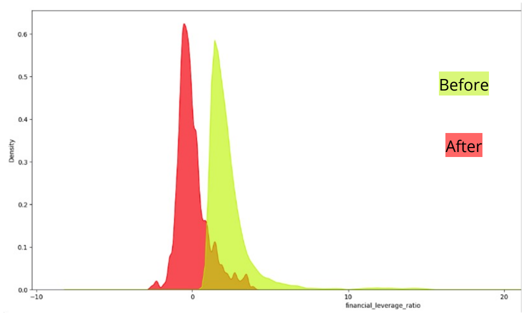
1. **Limit Extreme Values** by capping feature values at 3 std from the mean
2. **Interpolate Missing Values** with Industry Mean
3. **Feature Bias Reduction** using OLS regression to extract residuals and eliminate bias introduced by <u>company size</u> and <u>industry</u> differences.
4. **Z-Score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



Before



After



The model is in much better shape to focus on the true signals within preprocessed features

Methodologies

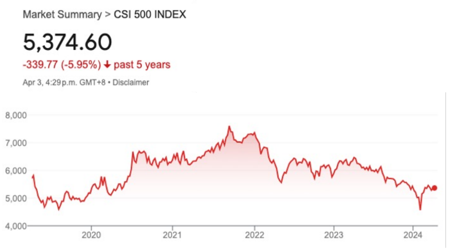
Algorithms:

1. **Naive Bayes Classifier -** simple probabilistic estimation
2. **XGBoost Classifier -** 5-fold cross-validation (shallow tree structure works best)

Evaluation:

1. **F1-Score Over time**
    a. Harmonic mean of precision & recall
    b. Identify potential degradation in performance over time

1. **Benchmark Comparison (CSI 500 INDEX)**
    a. Cumulative return %
    b. Maximum Drawdown %
    c. Proportions of months beating benchmark
    d. Information Ratio (excess return/standard deviation)

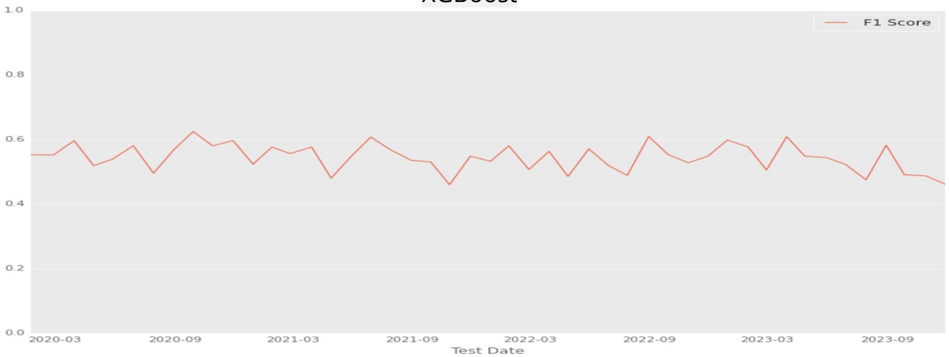


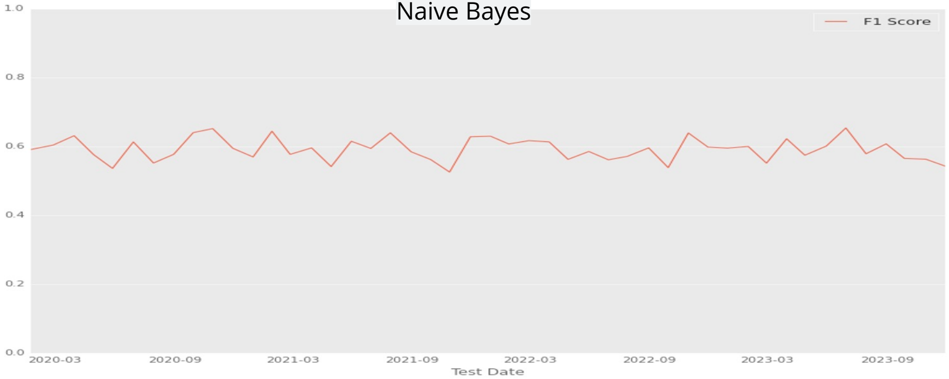
1. **Rank-based Backtesting**
    a. Segmented the portfolio into quantiles based on the model's predicted probabilities. (eg: does the top 10% quantile in probability perform better than the 40-50% quantile?)

Evaluation

F1-Score over time 2020-2023

XGBoost





- **F1 score:** harmonic mean of the precision and recall
- oscillates between 0.5 to 0.6
- no evident declining trend in performance in both models

Evaluation

Benchmark Comparison - XGBoost (Jan 2020 - Dec 2023 snapshot)

Metrics	Top 50%	Bottom 50%	Benchmark (CSI 500)	definition
Cumulative Return	**43.2%**	25.0%	1.0%	Jan 2020 - Dec 2023 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	$\text{Information Ratio} = \frac{E(R_i - R_b)}{\sigma_{ib}}$

* Segmented the XGBoost portfolio into quantiles based on the **model's confidence**
* **Top 50%** confidence within the subset of "**good stocks**"
* **Bottom 50%** confidence within the subset of "**good stocks**"
* {'learning_rate': 0.2, 'max_depth': 1, 'n_estimators': 10}.

Insights:

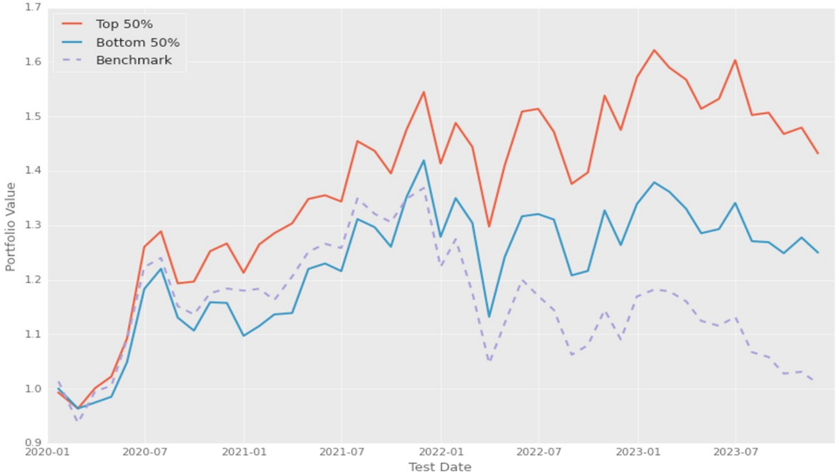
- **+42%** excess return compared to CSI 500
- **- 10%** in maximum loss of portfolio value
- The odds of beating the index is **61%**

CSI 500 INDEX

Evaluation

Rank-based Backtesting - XGBoost (Jan 2020 - Dec 2023 snapshot)

Metrics	Top 50%	Bottom 50%	Benchmark (CSI 500)	definition
Cumulative Return	**43.2%**	25.0%	1.0%	Jan 2020 - Dec 2023 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	$\text{Information Ratio} = \frac{E(R_i - R_b)}{\sigma_{ib}}$



* Segmented the XGBoost portfolio into quantiles based on the **model's confidence**
* **Top 50%** confidence within the subset of "**good stocks**"
* **Bottom 50%** confidence within the subset of "**good stocks**"
{'learning_rate': 0.2, 'max_depth': 1, 'n_estimators': 10}.

Insights:

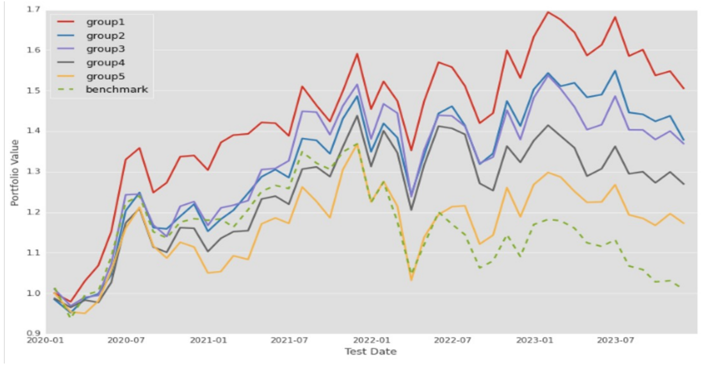
- top 50% quantile consistently outperforming the bottom quantile
- the model is **effective in distinguishing between stocks** with higher and lower potential performance
- Demonstrated enhanced resilience during market downturns

Evaluation

Rank-based Backtesting - XGBoost (Jan 2020 - Dec 2023 snapshot)

	Quantiles based on Confidence						
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 %
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	$\text{Information Ratio} = \frac{E(R_i - R_b)}{\sigma_{ib}}$

* g1 = Top 20 % predicted probability within the subset of "good stocks"
* g2 = Top 20 % - 40% predicted probability within the subset of "good stocks"
{'learning_rate': 0.2, 'max_depth': 1, 'n_estimators': 10}.



Insights:

- Higher confidence has better risk - control (less max losses)
- Result is consistent with previous breakdowns
- **G1 = Top 20%** predicted probability among the subset of "good stocks"

Conclusion & Final Remarks

- **Eliminated biases** introduced by company size and industry differences

- **Outperformed benchmark** in cumulative return % and maximum loss %

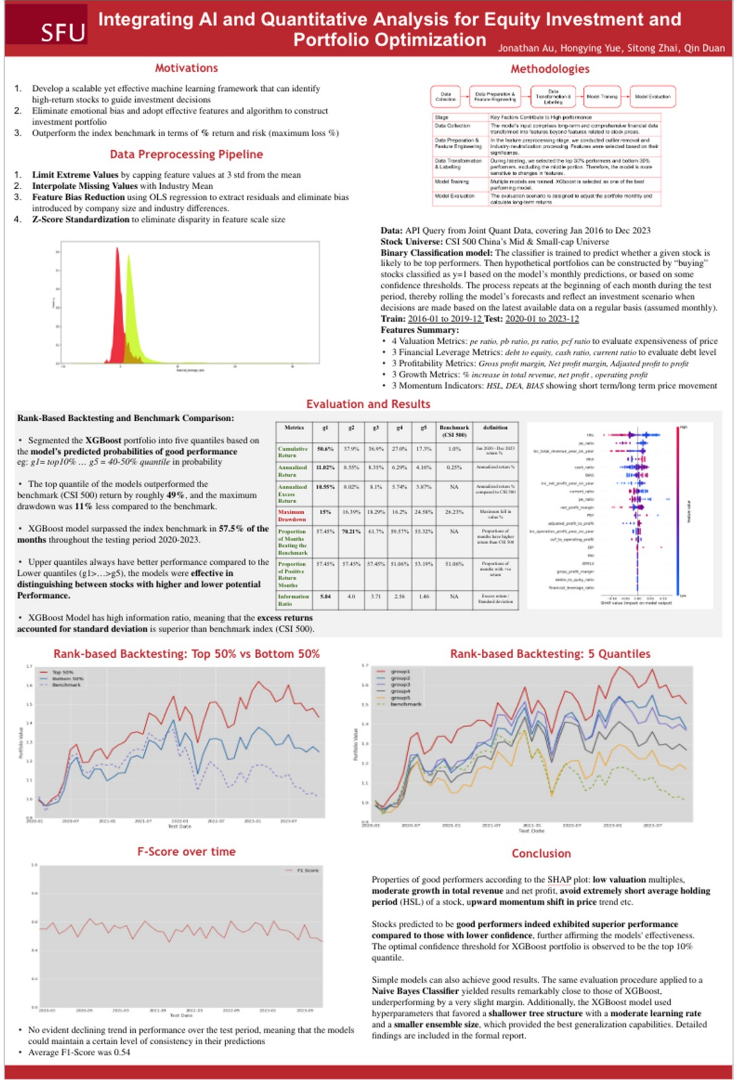
- Naive Bayes Classifier performed **similarity** to XGBoost Classifier with shallow and simple tree structure

- Identified attributes for good performing stocks
    a. **Low valuation** metrics (eg: low ps_ratio, pe_ratio)
    b. **Moderate growth** in total revenue and net profit (extremely high growth impacts negatively)
    c. **Long average holding** period of a stock over the past week.
    d. **Upward momentum** shift in price trend (eg: High DEA, BIAS)

- Ideas for Future Works
    a. Train / Test for longer horizons (eg: 2005-2024)
    b. Test if valid in other stock composites (*CSI 300 / S&P 500 / Hang Seng Index* etc)
    c. Explore for more features

Thank you

Shung Ho Jonathan Au:sha315@sfu.ca
Hongying Yue, hya134@sfu.ca
Sitong Zhai, sza210@sfu.ca
Qin Duan, qda18@sfu.ca



Integrating AI and Quantitative Analysis for Equity Investment and Portfolio Optimization

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 & Labeling → 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 & Labeling	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: 2016-01 to 2019-12 **Test:** 2020-01 to 2023-12
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.**
- XGBoost Model has high information ratio, meaning that the **excess returns accounted for standard deviation** is superior than benchmark index (CSI 500).

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

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, **upward 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.

Appendix 1

Feature Summary

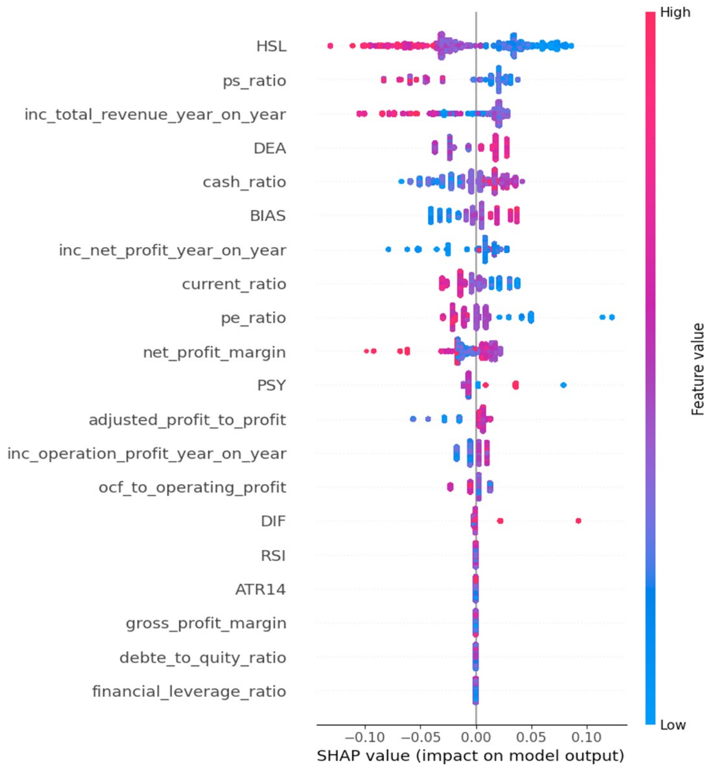
Valuation Metrics: These metrics assess how expensive a stock is relative to various financial fundamentals, reflecting market expectations for a company's growth and profitability.
PE ratio: Stock price to Earnings. Reflects how much investors are willing to pay per dollar of earnings
PB ratio: Stock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

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

Profitability Metrics: Higher profitability suggests a company is efficient in converting sales into actual profits.
Gross Profit Margin: (Sales Revenue - raw material cost) / Sales Revenue
Net Profit Margin: (Sales Revenue - raw material cost - operating cost) / Sales Revenue
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)

Growth: Measures the percentage growth in profit over time
Inc_total_revenue: Measures % increase in revenue
Inc_net_profit: Measures % increase in net profit
Inc_operating_profit: Measures % increase in operating profit

Momentum: Captures trading activities that measure short-term stock price movement relative to average to identify momentum shifts.
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.
DEA: Moving average difference between short-term price trend (10 days) versus long-term price trend (30 days), over 15 days.
BIAS: Compares the current stock price to its average over 20-day average to identify deviations from the typical level.



Appendix 2 - poster

Integrating AI and Quantitative Analysis for Equity Investment and Portfolio Optimization

SFU

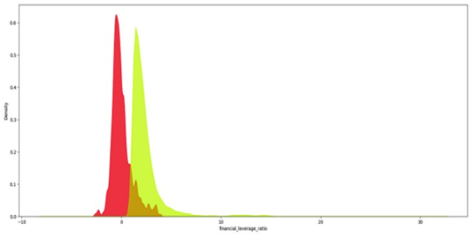
Jonathan Au, Hongying Yue, Sitong Zhai, Qin Duan

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: 2016-01 to 2019-12 **Test:** 2020-01 to 2023-12
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

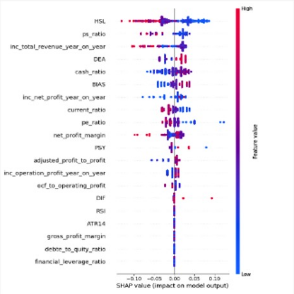
- The top quantile of the models outperformed the benchmark (CSI 500) return by roughly **49%**, and the maximum drawdown was **11%** less compared to the benchmark.

- XGBoost model surpassed the index benchmark in **57.5% of the months** throughout the testing period 2020-2023.

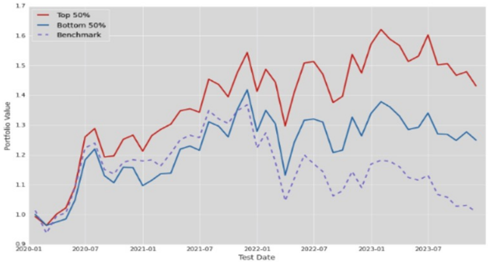
- Upper quantiles always have better performance compared to the Lower quantiles (g1>…>g5), the models were **effective in distinguishing between stocks with higher and lower potential Performance.**

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

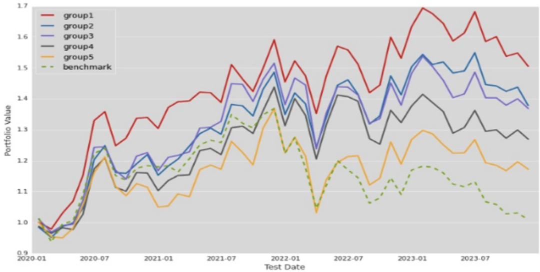
Metrics	g1	g2	g3	g4	g5	Benchmark (CSI 500)	definition
Cumulative Return	**50.6%**	37.9%	36.9%	27.0%	17.3%	1.0%	Jan 2020 - Dec 2023 return %
Annualized Return	**11.02%**	8.55%	8.35%	6.29%	4.16%	0.25%	Annualized return %
Annualized Excess Return	**10.55%**	8.02%	8.1%	5.74%	3.87%	NA	Annualized return % 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



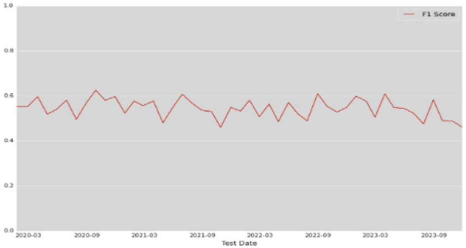
Rank-based Backtesting: Top 50% vs Bottom 50%



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, **upward momentum shift in price** trend etc.

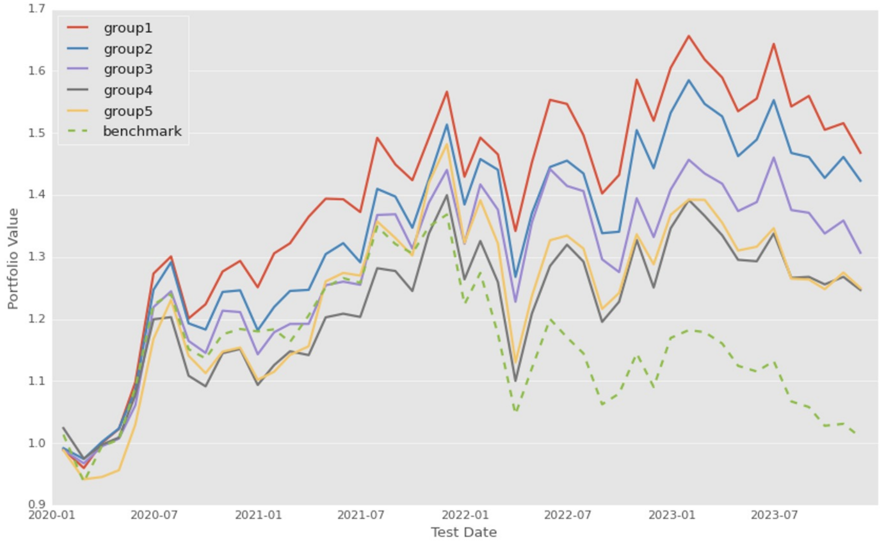
Stocks predicted to be **good performers indeed exhibited superior performance compared to those with lower confidence**, further affirming the models' effectiveness. The optimal confidence threshold for XGBoost portfolio is observed to be the top 10% quantile.

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.

Appendix 3

Naive Bayes Deep Dive

	G1	G2	G3	G4	G5	Benchmark
Cumulative Return	46.8%	42.3%	30.7%	24.7%	25.0%	1.0%
Annualized Return	10.3%	9.42%	7.07%	5.8%	5.86%	0.25%
Maximum Drawdown	14.33%	16.21%	14.75%	21.39%	23.75%	26.23%
Sharpe Ratio	1.23	1.03	0.59	0.35	0.34	-0.75
Annualized Excess Return	9.81%	8.93%	6.6%	5.4%	5.63%	0.0%
Monthly Maximum Excess Return	5.9%	6.53%	5.36%	5.15%	5.7%	0.0%
Proportion of Months Beating the Benchmark	57.45%	65.96%	63.83%	61.7%	59.57%	0.0%
Proportion of Positive Return Months	55.32%	55.32%	51.06%	55.32%	55.32%	51.06%
Information Ratio	4.82	3.68	2.93	2.74	2.57	NaN



Appendix 4

CSI 500 INDEX deep dive