Project Proposal: Integrating AI and Quantitative Analysis for Equity Investment

Team(4p): Shung Ho (Jonathan) Au, Hongying Yue, Sitong Zhai, Qin Duan

1. Research questions

List 3 questions that you intend to answer (1 point)

1. How can machine learning techniques be utilized to evaluate and classify best-performing stocks, enabling investors to make data-driven selections and achieve excess returns?
2. What financial indicators can serve as features to construct multifactor models for predicting stock returns? How can the choice of features be substantiated by economic theory?
3. How do the returns of portfolios constructed based on these models compare to benchmarks such as the CSI 500 Index, random stock selection etc.

2. Dataset utilization

List all the datasets you intend to use (1 point)

JQData Open Source: This research utilizes the JQData API, an open-source comprehensive quantitative financial data platform, to access and analyze extensive financial datasets. JQData offers access to a wide array of financial information, enabling users to review and compute various financial metrics.

- We will query CSI 500 Index composite stock data and company fundamentals from 2016 to 2023. The composite stocks comprise 500 small to mid-cap stocks traded on the Shanghai and Shenzhen stock exchanges.
- We will construct a dataset that includes the monthly return of each stock calculated from price data, along with valuation metrics, profitability metrics, momentum indicators etc. Each metrics category will involve multiple indicators, computed based on raw financial data. The detailed metrics to be used will be determined shortly.

3. Methodology

Give us a rough idea of how you plan to use the datasets to answer these questions. (2 points)

- Data Collection: Our study focuses on querying composite stock data from the CSI500 Index, which represents 500 mid and small stocks by market capitalization on the Shanghai and Shenzhen stock exchanges. In addition to stock price over time, we would also have access to company fundamentals covering the years 2016 to 2023. This includes income statements, balance sheets, and cash flow statements of companies listed in the CSI 500 Index.

- Data Exploration: Feature engineering in quantitative investing requires an understanding of financial information and metrics that may be used as predictive indicators of stock performance. We need to conduct multiple rounds of data exploration to identify/formulate “growth” related factors, “momentum” related factors, “valuation” related factors, “capital structure" related factors etc and other possible metrics for feature engineering, which we have to transform and compute from raw financial data at the data transformation stage.

- Data Cleaning:
  1. Financial data always contains errors such as extreme values / null values, we may need to handle those exceptions with industry mean.
  2. Also, since different "factors" may have vastly different scales and units, it is likely that we also have to normalize the data to ensure that all input features are on a comparable scale.
  3. To ensure stocks across different sectors are being compared on a like-for-like basis, we may have to isolate the pure effect of a factor by removing biases that could be introduced by external variables like industry classification and market capitalization etc. In simple words, we want to avoid sector-specific booms influencing the universal effectiveness of a factor across all industries.

- Data Integration: Although we query our financial and stock data from JQData API, we need to integrate data from multiple sources and combine them into one for training and testing. For example, daily/monthly price data for individual stocks, trailing data in profitability and company fundamentals, momentum indicators from trading etc.

- Data Analysis:
  1. To train our supervised learning model that classifies stocks on a monthly basis, we will label the top and bottom-performing stocks based on their monthly return. The rationale behind this approach is to train machine learning models to predict which stocks are likely to be top performers and bottom performers and to classify/rank stocks according to their likelihood of high return. We could then automatically construct a monthly investment portfolio with the top x% of stocks with the highest predicted scores.
  2. The analysis could involve the choice between machine learning algorithms, labelling techniques, the effectiveness of data cleaning, feature importance, the statistical difference between the portfolio constructed in different quantiles based on the likelihood of high returns etc. We will determine as we progress. The core analysis would be the comparison between the model and index benchmark (CSI500 INDEX), or random stock selection.
  3. To evaluate portfolio performance, we will use the first 4 years of monthly data as training and the subsequent 4 years as testing. In other words, we will start by training the model using the data from month 1 until 48 months, then use the model to test and simulate stock picks on month 49, and then on 50 until the end (96). Finally, we will plot the return (% change in the price of the portfolio) and compare it to an index benchmark or random stock selection.
  4. To evaluate the model performance over time, we will focus on some of the metrics used in the investment industry. Including "cumulative return", "Annualized return", "Maximum drawdown", "Percentage of month outperforming benchmark", and "Information Ratio" (excess return relative to the benchmark divided by the standard deviation).

- Data Product:
  1. A framework that utilizes machine learning models as a starting point to construct an equity portfolio.
  2. A report that summarizes the performance of our concept.

4. Expected impact

Think about that once your project is complete, what impacts it can make. Pick up the greatest one and write it down. (1 point)

Acknowledging the truth that approximately 90% of investment strategies underperform their benchmarks, the greatest impact of this project is not to promise a model that can consistently beat the market, but to rigorously examine whether a data-driven mindset can improve the efficiency of capital allocation.

This approach aims to contribute to a more efficient financial market, where investment decisions are informed by data-driven insights rather than speculative evidence, thereby offering a strategic advantage in navigating the complex dynamics of investing.

5. Potential challenges

Identify any anticipated obstacles and how you plan to address them. (1 point)

Financial datasets may contain errors, missing values, or inconsistencies. Implement rigorous data cleaning and preprocessing steps. We will try to address this by leveraging reputable financial data providers.

Financial markets are influenced by countless factors, including economic indicators, political events, and investor sentiment, making them inherently noisy and unpredictable. Properly benchmarking model performance and setting realistic expectations is crucial. We will not treat beating the benchmark as the only goal but use a variety of performance metrics to understand the model better.