2023 Lazard Data Science Technical Test

Predicting stock returns using news headline and sentiment.

Project Description

Our objective in this project is to predict monthly stock price returns of 25 selected US consumer companies using a dataset of news headline and sentiment scores. Candidates are expected to derive useful features from the headline texts and use those features to predict monthly stock returns along with various score factors in the news dataset.

We provide two objective variables, current-month returns and next-month returns, and the candidate should build models for each respectively. We also provide an optional monthly dataset of these companies' financial metrics as additional features to include in your model.

Project Deadline

The deadline is **Thursday, November 9, 2023 11:00 AM ET**. Please create a private GitHub repository with all relevant files and add **Haotian Zhang** (haotian.zhang@lazard.com) as a collaborator. Please email Haotian to confirm you've submitted the assignment.

Data Description

Brief descriptions of the datasets below:

- **consumer_comp_news_sentiment.csv**
  Includes the relevant news of the 25 companies from 01/2020 to 10/2023. Key columns include (not limited to) the following:
  * `TIMESTAMP_UTC`: publish date and time of the article
  * `RP_ENTITY_ID`: company identifiers for news dataset
  * `ENTITY_NAME`: company name
  * `RELEVANCE`: news article relevance to the company
  * `TOPIC`, `GROUP`, `TYPE`, `SUB_TYPE`: category of news event with increasing granularity
  * `EVENT_RELEVANCE`: news event relevance to the company
  * `EVENT_SENTIMENT_SCORE`: news event sentiment score
  * `EVENT_TEXT`: news event description
  * `HEADLINE`: news article headline

- **y_series_return.h5**
  Includes two objective series to predict-monthly price returns and 1-month forward price returns of the companies.
  * `security`: company tickers (or identifiers)
  * `date`: month-end date.
  * `MONTHLY_RETURN` refers to the stock price return (i.e. %change) in the *current* month ending with the date in `date` column.
  * `MONTHLY_RETURN_F1` is the stock price return in the month *after* the date in `date` column.

  Note that not all companies have a full history of price returns. For example, companies that go public after 01/2020 would have missing returns before the IPO month. Therefore, please exclude months with missing returns in your models.

- **x_df_comp_factors.h5**
  Includes financial metrics for a comprehensive view of a company's financial health, performance, and risk profile. From earnings data like "TRAIL_12M_NET_SALES" and "TRAIL_12M_EPS" to valuation ratios like "EV_TO_T12M_EBITDA" and risk measures like "ALTMAN_Z_SCORE", these factors are essential for making investment decisions. Note, the `date` column is the month-end date. It's optional to include these factors in your prediction model.

- **rp_full_ticker_mapping.xlsx**
  Includes the ID mapping (`RP_ENTITY_ID` <-> `security`) between the news dataset and other tables.

Final Deliverables

The final deliverable should consist of source code/notebooks to process raw datasets and build models. Final results including model output, evaluations, and visualizations – this can either be embedded in the final notebook, or in a separate write-up. In both cases, please include a brief discussion or interpretation of the overall research process and model performance. Please highlight any significant findings, factors, or models.

Evaluations

We will assess both the quality of the final output - including novelty and complexity of modeling, performance of the models, visualizations, and quality of analysis or discussion of the model – and the code quality of the scripts.

Bonus (Optional)

Retail investors and social media has an impact on market movements, and this can be gauged using web scraping. As a bonus, write a simple script to scrape online posts, extract features and use them to complement your model to predict returns. You can use any available libraries or projects you find online to facilitate extracting the posts.

Data Policy

All the data included in the project folder are for test purpose only, and they should be deleted from the candidate's local device(s) after completion and submission of the test. Any sharing and publishing of the data and project description, including uploading to public online repository and file drives, are strictly prohibited without permission from Lazard.

Reference Policy & Code of Conduct

We encourage the implementation of existing literature and use of any open-source libraries, as long as references to literature and code repositories are provided in your final deliverable. However, any answers or scripts specific to this project must be your own work. You may not share your solutions with anyone else. This includes anything written by you, as well as any official solutions if provided by Lazard. You may not engage in any other activities that will dishonestly improve your results or dishonestly improve or damage the results of others.

Questions

If you have any questions or issues related to data sets or project descriptions, please feel free to reach out to Haotian Zhang (haotian.zhang@lazard.com).