Financial Modelling & Testing

Earnings Prediction

Rohitashwa Chakraborty

Problem Statement

The objective of this project is to forecast earnings per share of publicly companies listed publicity over the next fiscal year. Should we be successful in our endeavor, it will allow us to create zero cost long-short portfolios on this strategy, which can then the back tested to check for profitability. Currently, all predictive models use financial data and indicators from the income statements and balance sheets for their analysis. However, in order to improve the quality of our predictions, we might want to consider other market and economic data including returns, inflation, interest rates as well as unstructed data which we will get from 10Ks and 10Qs filings, call transcripts etc.

Why use Unstructured Data?

Previously, we focused on predicting change in earnings per share (eps) by considering financial data exclusively. We plan to build on top of our models in the second iteration to augment our datasets with information extracted from sources of unstructured data. Some of these sources are SEC filings reports, like 10K(s) and 10Q(s), transcripts of call or public events etc.

Using unstructured data adds a new dimension to our analysis as it allows us to capture richer and high level information like sentiments, tone, confidence in the statements etc. Furthermore, we can skim through such data sources to look for key words which will let us understand the future plans for the company. Since financial statements are just a report on the history of the company, the numbers on this report can never capture the aforementioned information.

While text is the most common form of unstructured data, and we can extract sentiment, language complexity, look for words etc., it is only a small part of the big picture. Advanced techniques could include understanding the tone, detecting sarcasm and measuring confidence by analyzing the spectrogram from audio recordings. As is the case with every opportunity, this too comes full of its own challenges such as quality of recording, susceptibility to ambient noise, distinguishing between different voices etc.

Performing a cross-sectional analysis, across the market will give us better insights into what we might expect from the company in the following years than just using financial data.

Data Sources

The following would be our data sources for the financial indicators:

- Compustat (from WRDS database)
- treasury.gov (Data on economic health like interest on t-bills)
- FRED economic data
- Kenneth R French Data Library (Fama French factors to help build portfolios)

We can obtain unstructured data from the following sources:

- SEC Filings (scrape 10K(s), 10Q(s) and other reports made available by EDGAR)

For now, we could use this as our only source of unstructured data since the Meta team has already made available all of the past 10K(s) and 10Q(s) in the google drive.

As done previously in the fallen angels assignment, we must process the text files to quantify the different aspects like sentiment, complexity (fog score) etc., and this information is appended to our financial indicators dataset. One must note that the preprocessing step is computationally extremely intensive and we are often limited by the power (RAM) of the computers at our disposal. Such a bottle neck makes it impossible for us to look too far back in time and analyze the trend of sentiments or complexity (for example) over time.

The text could include the following:

- Sentiment Analysis
- Fog Index
- Frequency of the 50 most risky financial words from Harvard List.
- Bi Grams (if it is computationally not limiting).
- Document similarity score.

Creating Sorted Portfolio

For this step, we will use the eps predictions from our model and sort the companies each year into five quantiles; from the top performers to the bottom.

Should we see a statistically significant and monotonous change in eps over the 5 bins, we can conclude that building long-short portfolios on predicted eps change is a marketable strategy indeed.

In the interest of stability, we might choose to ignore the quantile 1 and 5 (top and bottom most companies) since they might contain a log of outliers and consider long/shorting quantile 2 &4 instead. As in the norm while creating zero-sum portfolios, we short the bottom performers and long the top performers by an equal amount, therefore, hedging our risks. The portfolio will be reconstituted yearly, however, we might choose to rebalance it monthly.

If this Portfolio has a significant and high positive Alpha, compared to market or the market index with the most similar risk and exposure, we will conclude that this is a marketable trading strategy, and we would invest more in improving the performance of the model.

Limitations

From our past experience, especially during the prediction of fallen angels, we observed that the main issue with such assignment is the analysis of text data. The scraping itself is a tedious and often not 100% reliable (due to network errors, request timeouts etc.). Furthermore, since the reports are very large, Text analysis is excruciatingly slow, with the sentiment analysis. alone often taking up to 7 hours to run. Collab is unreliable and local machines are often not able to process the whole data frame either .

To help with this we might limit our analysis to just 1 or 2 industry sectors.

References

- https://zicklin.baruch.cuny.edu/wp-content/uploads/sites/10/2019/12/Improving-Earnings-Predictions-with-Machine-Learning-Hunt-Myers-Myers.pdf
- https://www.marketwatch.com/story/how-oil-traders-are-using-satellites-to-keep-an-eye-on-an-increasingly-unpredictable-market-2019-10-04