STOCK PRICE PREDICTIONS USING 10-K FILINGS

In this case study, we will apply predictive analytics in a case study to predict stock prices of companies by using their public 10-K filings.

Background: 10-k Filings of Public Companies

The 10-k filing is a yearly summary report that all public companies must register with the SEC to report on various aspects, including current finances. It includes various information sets, including but not limited to risk factors, market dynamics and managerial reports, which can significantly affect investors' opinions. With proper analysis, the 10-k analysis can facilitate investors' selection process for their portfolios, and reduce non-systemic risks while maximizing returns. Can applied predictive analytics on a corpus 10-k filings be used to deliver superior financial returns?

Challenges and Opportunity

There are at least 20,000 words in each 10-k file, and it takes tremendous efforts for investors to go through and digest each 10-k filing of all public-listed companies. This provides a golden opportunity to use text analytics. We can potentially utilize modern machine learning and natural language processing techniques to extract and summarize valuable information from 10-k files without manually reading them, and 'combine' this information with chosen stock price features (such as the stock price time series of the company) to build a prediction model. Ideally, the model can use the current year 10-k files contents and stock price features (current and past) to make a prediction of the company's average stock price in the next year.

Publicly available dataset

While 10-k filings are publicly available, they can be difficult to download in one place. Luckily, there is a publicly available source already for a limited subset of companies. Specifically, 10-K reports from thousands of publicly traded U.S. companies, published in 1996–2006 and stock return volatility measurements in the twelve-month period before and the twelve-month period after each report, are available (version 1.0 released March 31, 2009) at http://www.cs.cmu.edu/~ark/10K/ at the time of writing. In addition to this data, we can also download (from Kaggle or other sources) the stock price time series of some of these companies.

Data Preprocessing

When downloading two or more datasets, an important step is to 'join' the datasets by knowing how identifiers in one dataset link to identifiers in another dataset. In this case, you would want to use the 'stock ticker' symbol as the identifier, although other features could also potentially be used. If necessary, entity resolution algorithms such as you studied in class could be employed. For the text files, since each 10-k file contains 20,000 ~ 40,000 words and characters, which includes both informative and meaningless portions, the text should be further processed before vectorizing and feature extraction. An easy preprocessing pipeline is to first exclude special characters, and then eliminate words that are too short or that are stop-words. Indeed, we find that even these simple preprocessing steps can reduce file size by around 50%. Other preprocessing steps can also be applied (see case questions).

Model building

Once the data is prepared, we need to build a predictive model. In this case, text features and 'past' stock prices comprise our potential set of features, and the current stock price is what we are trying to predict. You must be very careful in using *past* data to predict *current* stock prices. If your model works, then by extension, you could use *present* data to predict *future* stock prices.

Case Questions

i) Suppose you were told that you could use either the previous periods' stock prices or the current period's 10-K filings (text) but not both. Discuss how your approach to the problem would differ and what the main kinds of techniques are that you would consider for each.

ii) For the two scenarios mentioned above, provide at least three reasonable baselines each.

iii) What is your understanding of the difference between a *joint* model and an *ensemble* model? If you had to 'combine' all the baselines in (ii), which paradigm typically applies?

iv) What makes data pre-processing of 10-K more challenging (potentially) than a regular text document? Try to find and download a few 10-K filings, and be concrete in your answers.

v) Expand upon the data preprocessing section in the case, including how you would deal with the stock price portion of the data. For example, what steps would you take to deal with outliers and how would you define an outlier? Do the stock prices need to be normalized in some way or should you use the raw values? *Hint: Don't forget the role of inflation, and possibly other factors.*

vi) Suppose the date is 1 January 2006, and you had 10-k filings from 2000-2005, and stock prices all the way till your current date. You limit yourself to companies that were on public exchanges during this time (i.e. if any companies went bankrupt or were listed on the stock exchange during this time, you exclude them from your data). Describe in detail how you would set up training, validation and testing for your models. Specifically, what data would be used for training and how? What are some options for seeing if your model is working? And what kind of performance would you like to see before you use the model to make investing decisions for you?