Data Description and Variable Definitions

Forecasting Earnings Project

# Raw Data

The primary data source is the Compustat database. This database contains one row per firm per year. The columns contain the firm’s annual accounting numbers. The secondary data source is the CRSP database.[[1]](#footnote-1) CRSP contains daily stock price information for firms.

In the initial stages of the project, we will only use Compustat data. However, we need to filter the Compustat data based on CRSP variables. For example, we only want firms whose stocks are traded on certain exchanges, and firms whose stock price exceeds $1. Such information is in CRSP and must therefore be merged with Compustat.

I download the raw data from [WRDS](https://wrds-www.wharton.upenn.edu/) (Wharton Research Data Services).[[2]](#footnote-2) I have an account through our university. My username is *vanand* and my password is *Badri929\_1911*.

# Final Data Sample

I attempted to replicate the sample in the Cao and You (2020) paper that you read. Since WRDS has historically been SAS-based, I downloaded the data in SAS format and wrote SAS code to merge the Compustat and CRSP datasets. I will make the SAS code available to you in case you wish to validate it.

The final dataset is in the file CaoYouSample.parquet.[[3]](#footnote-3) I will make this available to you as well.

# Variable Definitions

The file CaoYouSample.parquet contains 71 columns. I define and describe each in the table below.

In the data file, variable names with suffix “\_F1” show 1-year ahead values. Thus, E is earnings, and E\_F1 is earnings in the following year. I included this to make it easy to train ML models as *E\_F1* will be the target.

The suffix “\_D1” indicates a first difference. Say variable X has value 10 in year 2000 and value 8 in 1999. Then X\_D1 in 2000 will be 10 – 8 = 2.

| Variable | Definition |
| --- | --- |
| Identifying Information | |
| GVKEY | Unique company identifier from the Compustat database |
| DATADATE | Date on which the row appeared in Compustat. This is usually the date on which the accounting information was released by the company to the public. |
| FYEAR | Fiscal year corresponding to the row |
| LPERMNO | Unique company identifier from the CRSP database |
| SIC | A 4-digit industry code. A complete listing can be found [here](https://siccode.com/). |
| CONM | Company name |
| TIC | Ticker symbol (symbol used by the stock exchange) |
| CUSIP | Identifier for the company’s common stock |
| FYR | Number indicating the company’s fiscal year end month. This will be from 1 – 12. |
| SHRCD | Share code |
| EXCHCD | Exchange code |
| FiscalYearEnd | Date on which fiscal year ends |
| FYEND\_plus\_3mos | Fiscal year end date plus 3 months |
| Balance sheet items | |
| ACT | Current assets |
| AP | Accounts payable |
| AT | Total assets |
| CEQ | Common equity |
| CHE | Cash and short-term investments |
| DLC | Current liabilities |
| DLTT | Long-term debt |
| INTAN | Intangible assets |
| INVT | Inventory |
| IVAO | Investment and Advances Other |
| LCT | Current liabilities |
| LT | Total liabilities |
| PPENT | Property, plant, and equipment |
| RECT | Receivables total |
| TXP | Income taxes payable |
| Income statement items | |
| COGS | Cost of goods sold |
| DP | Depreciation and amortization |
| NOPIO | Nonoperating Income (Expense) Other |
| SALE | Sales / revenue |
| TXT | Income taxes total |
| XAD | Advertising expense |
| XIDO | Extraordinary Items and Discontinued Operations |
| XINT | Interest and Related Expense |
| XRD | Research and development expense |
| XSGA | Selling, general, and administrative expense |
| E | Earnings. Computed as income before extraordinary items (IB) minus special items (SPI) |
| Other items | |
| CFO | Cash flow from operations. Computed as operating activities net cash flow (OANCF) minus extraordinary items and discontinued operations (XIDOC) |
| CSHO | Common shares outstanding |
| DVC | Dividends paid |

1. CRSP stands for the Center for Research in Security Prices at the University of Chicago. [↑](#footnote-ref-1)
2. Wharton is the business school at the University of Pennsylvania. [↑](#footnote-ref-2)
3. Parquet is my preferred format for storing tabular data. It is a columnar storage format and includes compression by default. Reading and writing Pandas data frames to Parquet is very fast. However, to use it, you must install the pyarrow package from Anaconda. [↑](#footnote-ref-3)