

**Data Description Sheet for
“Predicting Future Earnings Changes Using Machine Learning and Detailed Financial
Data”**

by Xi Chen, Yang Ha (Tony) Cho, Yiwei Dou, and Baruch Lev

1. *A description of which author(s) handled the data and conducted the analyses.*

Co-author Yang Ha (Tony) Cho handled the data and conducted the analyses.

2. *A detailed description of how the raw data were obtained or generated, including data sources, the specific date(s) on which data were downloaded or obtained, and the instrument used to generate the data (e.g., for surveys or experiments). We recommend that more than one author is able to vouch for the stated source of the raw data.*

- a. We downloaded XBRL filings submitted between June 15, 2012, and March 31, 2018, from the SEC website (<https://www.sec.gov/dera/data/financial-statement-data-sets.html>) in December 2019.
- b. We obtained firm-level financial data from Compustat and stock return data from CRSP. The data were downloaded in September 2020.
- c. We obtained pro forma earnings and analyst forecast data from I/B/E/S in September 2020.
- d. Transaction costs were estimated using a code downloaded from Joel Hasbrouck’s website (<http://people.stern.nyu.edu/jhasbrou/Research/GibbsCurrent/gibbsCurrentIndex.html>) in February 2020.
- e. We obtained Fama-French factors and industry classification from https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html in October 2020.
- f. We obtained XBRL taxonomy and presentation map from the XBRL US website (<https://www.fasb.org/cs/ContentServer?c=Page&cid=1176169699514&d=&pagename=FASB%2FPage%2FSectionPage>) in September 2020.

All co-authors can vouch for the stated source of the raw data.

3. *If the data are obtained from an organization on a proprietary basis, the authors should privately provide the editors with contact information for a representative of the organization who can confirm data were obtained by the authors. The editors would not make this information publicly available. The authors should also provide information to the editors about the data sharing agreement with the organization (e.g., non-disclosure agreements, any restrictions imposed by the organization on the authors, such as restrictions to publish certain results).*

All of the data were obtained from the public databases as described above.

4. *A complete description of the steps necessary to collect and process the data used in the final analyses reported in the paper. For experimental and survey papers, we require information*

about the instructions and instruments used to generate the data, subject eligibility and/or selection, as well as any exclusion criteria.

We followed the steps described below to convert the XBRL data to the firm-year level sample used to predict one-year-ahead earnings changes:

- 1) We obtained submission and numerical data from XBRL filings submitted between the second quarter of 2012 and the first quarter of 2018. We retained only the observations with a non-missing reporting period ending on or after June 15, 2012.
- 2) Using the accession number identifier, we merged the submission and numerical data. We kept only 10-K and 10-K/A and excluded any observations that are not fiscal year-end filings. We retained only numerical items that are identified with a U.S. GAAP tag and that are attributed to a consolidated entity (i.e., the “version” field starts with “us-gaap”; the “coreg” field is empty.)
- 3) We obtained pro forma earnings from I/B/E/S and kept only the pro forma earnings that differ from the U.S. GAAP earnings. We adjusted the pro forma earnings for a firm-specific trend by subtracting the average change in earnings over the past four years. We merged the pro forma earnings data with the data from Step (2).
- 4) We merged the data from Step 3) with a) firm-level financial data from Compustat for the construction of predictors in Ou and Penman (1989) and Nissim and Penman (2001), b) analyst forecasts from I/B/E/S, and c) stock price data from CRSP. We excluded observations without a stock price at the portfolio formation date.
- 5) We required that observations have non-missing and non-zero total assets in the XBRL filings.
- 6) Using the reporting period and filing date from the submission data, we required that observations be filed within three months following a fiscal year-end. Then, we kept only the most recent data (i.e., if a company has an XBRL 10-K submission and an XBRL 10-K/A submission to revise financial statements (but not footnotes) before the portfolio formation date, we merge the two submissions by using the revised financial statement items from the 10-K/A and the footnote items from the 10-K).
- 7) Using the XBRL taxonomy and the data from Step 6), we kept only those tags that are used at least once each year from 2012 to 2018.
- 8) For each firm-year observation from Step 7), we selected numerical data that correspond to current and lagged values by using the reporting period to identify current and lagged fiscal years. We divided each monetary item by its respective fiscal year’s total assets. For each numerical item, we created an annual percentage change variable. We retained only the current variable, lagged variable, and percentage changes.
- 9) The missing predictor values are all replaced with zeros.

We followed the step described below to categorize each tag into a financial statement category. Using the XBRL taxonomy and presentation map, we first prioritized assigning tags to a financial statement category (i.e., balance sheet, cash flow statement, income statement, comprehensive income statement, and stockholders’ equity statement). If a tag was not associated with a financial statement category, it was classified into footnotes. Whenever a tag was associated with multiple financial statement categories, it was placed into one of the financial statement categories with a more natural fit. Online Appendix Table A8 of the paper lists these instances and their respective financial statement categories.

5. *The computer programs or code used to convert the raw data into the final dataset used in the analysis plus a brief description that enables other researchers to use this program. The purpose of this requirement is to facilitate replication and to help other researchers understand in detail how the raw data were processed, the final sample was formed, variables were defined, outliers were treated, etc. This code or programming is in most circumstances not proprietary. However, we recognize that some parts of the code or data generation process may be proprietary, including from the authors' perspective. Therefore, instead of the code or program, researchers can provide a detailed step-by-step description of the code or the relevant parts of the code such that it enables other researchers to arrive at the same final dataset used in the analysis. In such cases, the authors should inform the editors upon initial submission, so that the editors can consider an exemption from the code sharing requirement.*

We used SAS to estimate transaction costs and R to convert the raw data into the final dataset. Due to computing power, R was used in conjunction with grid computers to conduct the analyses. In the primary analyses, R packages *gbm*, *randomForest*, and *caret* were used. Please see the code file (CCDL Code 20220212.pdf) and identifier file (adsh.txt) for details.

6. *An assurance that the data and programs will be maintained by at least one author (usually the corresponding author) for at least six years, consistent with National Science Foundation guidelines.*

Co-author Yang Ha (Tony) Cho will maintain the data and programs for at least six years.