CMPT 733 Machine learning to detect misstated financial statements

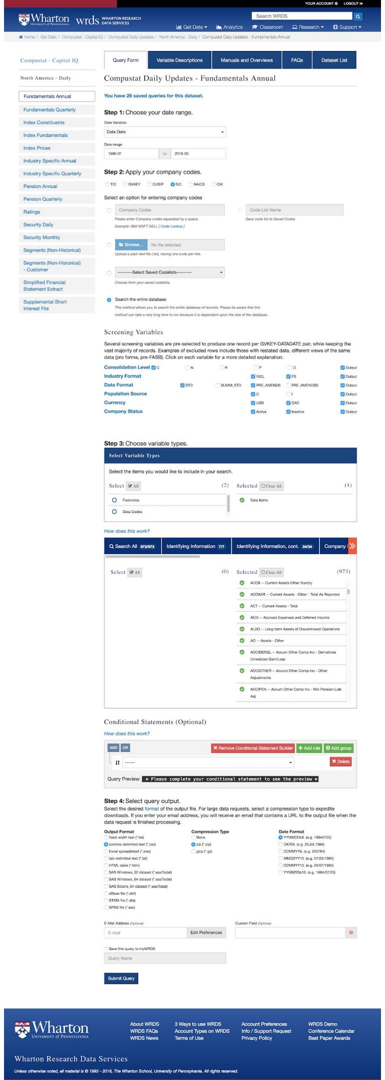
Katrina Ni, Leiling Tao

Motivation

- Financial statement fraud is deliberate misrepresentation, misstatement or omission of financial statement data, which accounts for 10% of white collar crimes (Association of Certified Fraud Examiners).
- Typically, financial statement contains thousands of fields which follow thousands of rules, presenting a daunting challenge for auditors to detect abnormal fields.
- We aim to automate the process of pre-screening potentially misstated financial statements by using machine learning techniques.

Progress Report

Data Collection	Data Preprocessing	EDA
Data source: WRDS - Wharton Research Data Services Raw data ranges from Jan 1980 to Mar 2018 with 435497 observations and 981 features	Exclude observations with missing or zero values for total assets or total equity (~15%) Feature selection: ● can identify restatement, ● have significant effect (control variables such as industry group), ● accounting terms (4 levels of importance, starting with the most important ones) drop features with > 50% missing values Extract reasons of restatement: fraud and error (~2.7%), merger and acquisition (~1.4%), change in accounting rules (~0.1%)	Extremely unbalanced classes => unsupervised learning is probably a better option Highly correlated accounting variables => Looking for dimensionality reduction Look into features that have significant effect according to prior research => Focus on a specific country (USA) or an industry group in a specific year first



YOUR ACCOUNT LOGOUT

Wharton UNIVERSITY of PENNSYLVANIA wrds WHARTON RESEARCH DATA SERVICES

Search WRDS

Get Data ▾ Analytics Classroom Research ▾ Support ▾

Home / Get Data / Compustat - Capital IQ / Compustat Daily Updates / North America - Daily / Compustat Daily Updates - Fundamentals Annual

Compustat - Capital IQ

North America - Daily

- Fundamentals Annual
- Fundamentals Quarterly
- Index Constituents
- Index Fundamentals
- Index Prices
- Industry Specific Annual
- Industry Specific Quarterly
- Pension Annual
- Pension Quarterly
- Ratings
- Security Daily
- Security Monthly
- Segments (Non-Historical)
- Segments (Non-Historical) - Customer
- Simplified Financial Statement Extract
- Supplemental Short Interest File

Query Form | Variable Descriptions | Manuals and Overviews | FAQs | Dataset List

Compustat Daily Updates - Fundamentals Annual

You have 28 saved queries for this dataset.

Step 1: Choose your date range.

Date Variable: Data Date

Date range: 1980-01 to 2018-03

Step 2: Apply your company codes.

○ TIC ○ GVKEY ○ CUSIP ● SIC ○ NAICS ○ CIK

Select an option for entering company codes

○ Company Codes
Please enter Company codes separated by a space.
Example: IBM MSFT DELL [Code Lookup]

□ Code List Name
Save code list to Saved Codes

○ Browse... No file selected
Upload a plain text file (.txt), having one code per line.

○ ---------Select Saved Codelists---------
Choose from your saved codelists.

● Search the entire database
This method allows you to search the entire database of records. Please be aware that this method can take a very long time to run because it is dependent upon the size of the database.

Screening Variables

Several screening variables are pre-selected to produce one record per GVKEY-DATADATE pair, while keeping the vast majority of records. Examples of excluded rows include those with restated data, different views of the same data (pro forma, pre-FASB). Click on each variable for a more detailed explanation.

Consolidation Level ☑ C	□ N	□ R	□ P	□ D	☑ Output
Industry Format			☑ INDL	☑ FS	☑ Output
Data Format	☑ STD	□ SUMM_STD	☑ PRE_AMENDS	□ PRE_AMENDSS	☑ Output
Population Source			☑ D	□ I	☑ Output
Currency			☑ USD	☑ CAD	☑ Output
Company Status			☑ Active	☑ Inactive	☑ Output

Step 3: Choose variable types.

Select Variable Types

Select the items you would like to include in your search.

Select ☑ All (2): ○ Footnotes ○ Data Codes

Selected □ Clear All (1): Data Items

How does this work?

Search All 973/973 | Identifying Information 7/7 | Identifying Information, cont. 34/34 | Company

Select ☑ All (0)

Selected □ Clear All (973)

- ACOX -- Current Assets Other Sundry
- ACOXAR -- Current Assets - Other - Total As Reported
- ACT -- Current Assets - Total
- AEDI -- Accrued Expenses and Deferred Income
- ALDO -- Long-term Assets of Discontinued Operations
- AO -- Assets - Other
- AOCIDERGL -- Accum Other Comp Inc - Derivatives Unrealized Gain/Loss
- AOCIOTHER -- Accum Other Comp Inc - Other Adjustments
- AOCIPEN -- Accum Other Comp Inc - Min Pension Liab Adj

Conditional Statements (Optional)

How does this work?

AND OR ✖ Remove Conditional Statement Builder ✚ Add rule ✚ Add group

------ ✖ Delete

Query Preview: `* Please complete your conditional statement to see the preview *`

Step 4: Select query output.

Select the desired format of the output file. For large data requests, select a compression type to expedite downloads. If you enter your email address, you will receive an email that contains a URL to the output file when the data request is finished processing.

Output Format
○ fixed-width text (*.txt)
● comma-delimited text (*.csv)
○ Excel spreadsheet (*.xlsx)
○ tab-delimited text (*.txt)
○ HTML table (*.htm)
○ SAS Windows_32 dataset (*.sas7bdat)
○ SAS Windows_64 dataset (*.sas7bdat)
○ SAS Solaris_64 dataset (*.sas7bdat)
○ dBase file (*.dbf)
○ STATA file (*.dta)
○ SPSS file (*.sav)

Compression Type
○ None
● zip (*.zip)
○ gzip (*.gz)

Date Format
● YYMMDDn8. (e.g. 19840725)
○ DATE9. (e.g. 25JUL1984)
○ DDMMYY6. (e.g. 250784)
○ MMDDYY10. (e.g. 07/25/1984)
○ DDMMYY10. (e.g. 25/07/1984)
○ YYMMDDs10. (e.g. 1984/07/25)

E-Mail Address (Optional): E-mail | Edit Preferences

Custom Field (Optional)

□ Save this query to myWRDS

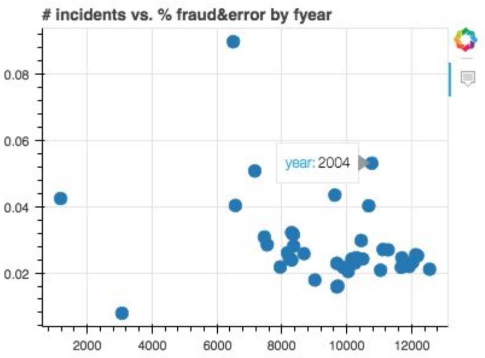
Query Name

Submit Query

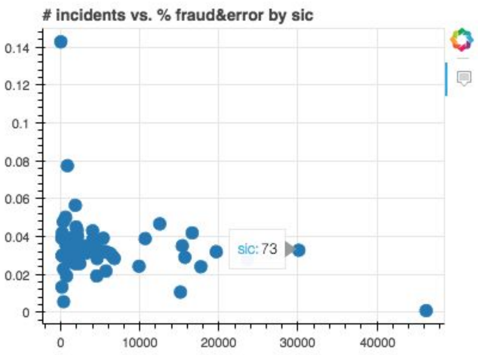
Wharton UNIVERSITY of PENNSYLVANIA

About WRDS | WRDS FAQs | WRDS News | 3 Ways to use WRDS | Account Types on WRDS | Terms of Use | Account Preferences | Info / Support Request | Privacy Policy | WRDS Demo | Conference Calendar | Best Paper Awards

Wharton Research Data Services

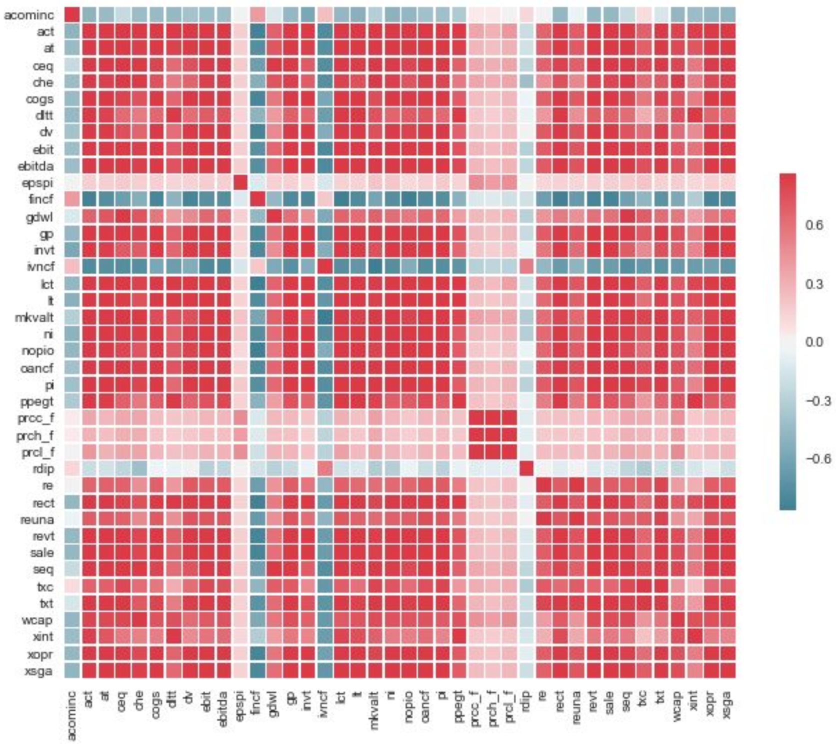


incidents vs. % fraud&error by fyear



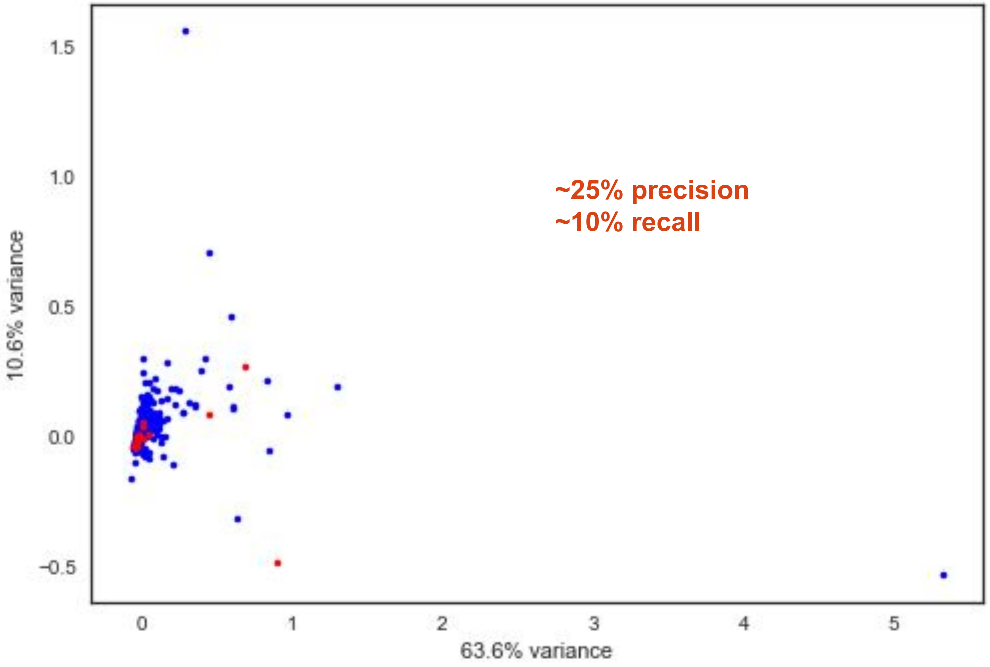
incidents vs. % fraud&error by sic

Need to focus on a subset of the data with reasonable number of observations and error/fraud rate
=> start with **USA, year 2004 and sic 73**



Most accounting features are highly correlated

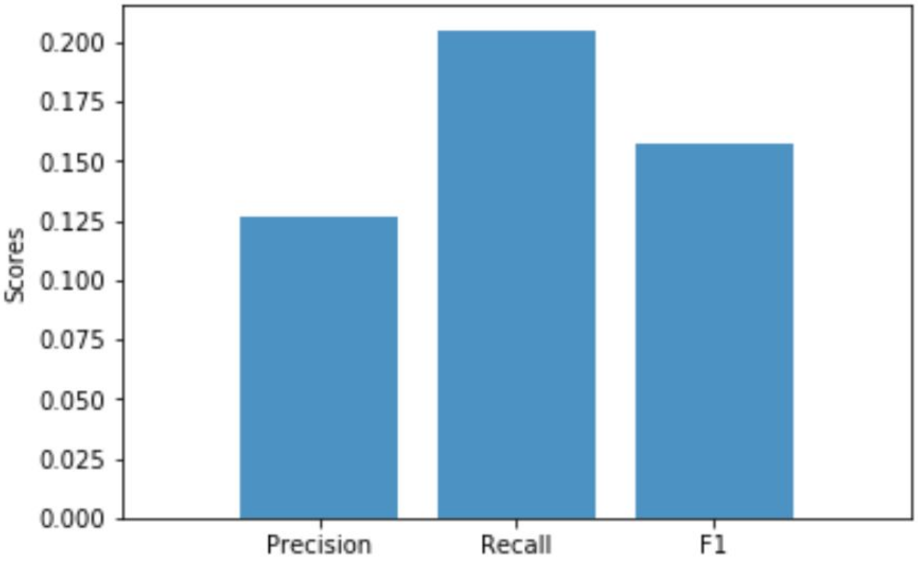
Initial **Dimensional Reduction** with USA, year 2004, industry code 73 (32 abnormalities among 831 data points)



~25% precision
~10% recall

Outlier detection using local outlier factor (LOF)

- Using data from 2014, industry code 73, for each variable, we calculated its percent change compared to 2013 (called horizontal comparison in accounting)
- Using these new features, we tried to detect outliers using local outlier factor



Abnormal changes in features are not good indicators of frauds!

Rather, we will focus on abnormal changes in relationships among variables in the future

Future Work

Unsupervised Learning	Supervised Learning	Neural network	External Dataset
Deploy several clustering and outlier detection methods: ● K-nearest-neighbors ● Isolation forest ● Time series analysis (ARIMA and LSTM)	Classification methods (need to consider imbalance): ● Random forest ● Gradient boosting ● SVM	Focus on relationships and patterns among variables, we will construct separate NN models to explore if and how do the structure of fraudulent statements differ from correct ones	Include industry specific indices to explore whether certain changes are more likely to result in frauds, such as ● Pension data ● Credit ratings ● Short interest ● Executive compensation