Corporate Fraud, LDA, and Econometrics

DSSG · 2019 March 27

Dr. Richard M. Crowley
SMU

rcrowley@smu.edu.sg · @prof_rmc

Slides: rmc.link/DSSG

The problem

> How can we *detect* if a firm is *currently* involved in a major instance of *misreporting*?

- *Detect*: Classification problem
- *Currently*: Prediction problem
- *Misreporting*: The accounting side
- The approach combines…

- Business insight
- Economic theory
- Psychology theory

- **Statistics**
- **Machine learning**
- **Careful econometrics**

Why do we care?

> The 10 most expensive US corporate frauds cost *shareholders* **12.85B USD**

- The above, based on Audit Analytics, ignores:
  - *GDP impacts*: Enron's collapse cost **~35B USD**
  - *Societal costs*: Lost jobs, economic confidence
  - Any *negative externalities*, e.g. compliance costs
  - *Inflation*: In current dollars it is even higher

> Catching even 1 more of these as they happen could save billions of dollars

What is Misreporting?

Misreporting: A simple definition

> Errors that affect firms' accounting statements or disclosures which were done seemingly *intentionally* by management or other employees at the firm.

Traditional accounting fraud

1. A company is underperforming
2. Management cooks up some scheme to increase earnings
   - Wells Fargo (2011-2018?)
     - Fake/duplicate customers and transactions
3. Create accounting statements using the fake information

Other accounting fraud types

- Dell (2002-2007)
  - *Cookie jar reserve* (secret payments by Intel of up to **76%** of quarterly income)
    1. The company is overperforming
    2. “Save up” excess performance for a rainy day
    3. Recognize revenue/earnings when needed to hit future targets
- Apple (2001)
  - *Options backdating*
- China North East Petroleum Holdings Limited
  - *Related party transactions* (transferring 59M USD from the firm to family members over 176 transactions)
- CVS (2000)
  - *Improper accounting treatments* (Not using mark-to-market accounting to fair value stuffed animal inventories)
- Countryland Wellness Resorts, Inc. (1997-2000)
  - Gold reserves were actually… dirt

Where are these disclosed? (US)

1. US SEC AAERs: Accounting and Auditing Enforcement Releases
   - Highlight larger/more important cases, written by the SEC
   - Example: The *Summary* section of this AAER against Sanofi
2. 10-K/A filings ("10-K" $\Rightarrow$ annual report, "/A" $\Rightarrow$ amendment)
   - Note: not all 10-K/A filings are caused by fraud!
     - Benign corrections or adjustments can also be filed as a 10-K/A
     - Note: Audit Analytics' write-up on this for 2017
3. By the US government through a 13(b) action
4. In a note inside a 10-K filing
   - These are sometimes referred to as "little r" restatements
5. In a press release, which is later filed with the US SEC as an 8-K
   - 8-Ks are filed for many other reasons too though

> Original disclosure motivated by management admission, government investigation, or shareholder lawsuit

Where are we at?

Fraud happens in many ways, for many reasons

- All of them are important to capture
- All of them affect accounting numbers differently
- None of the individual methods are frequent…

It is disclosed in many places. All have subtly different meanings and implications

- We need to be careful here (or check multiple sources)

This is a hard problem!

Predicting Fraud

Main question and approaches

> How can we *detect* if a firm is *currently* involved in a major instance of *misreporting*?

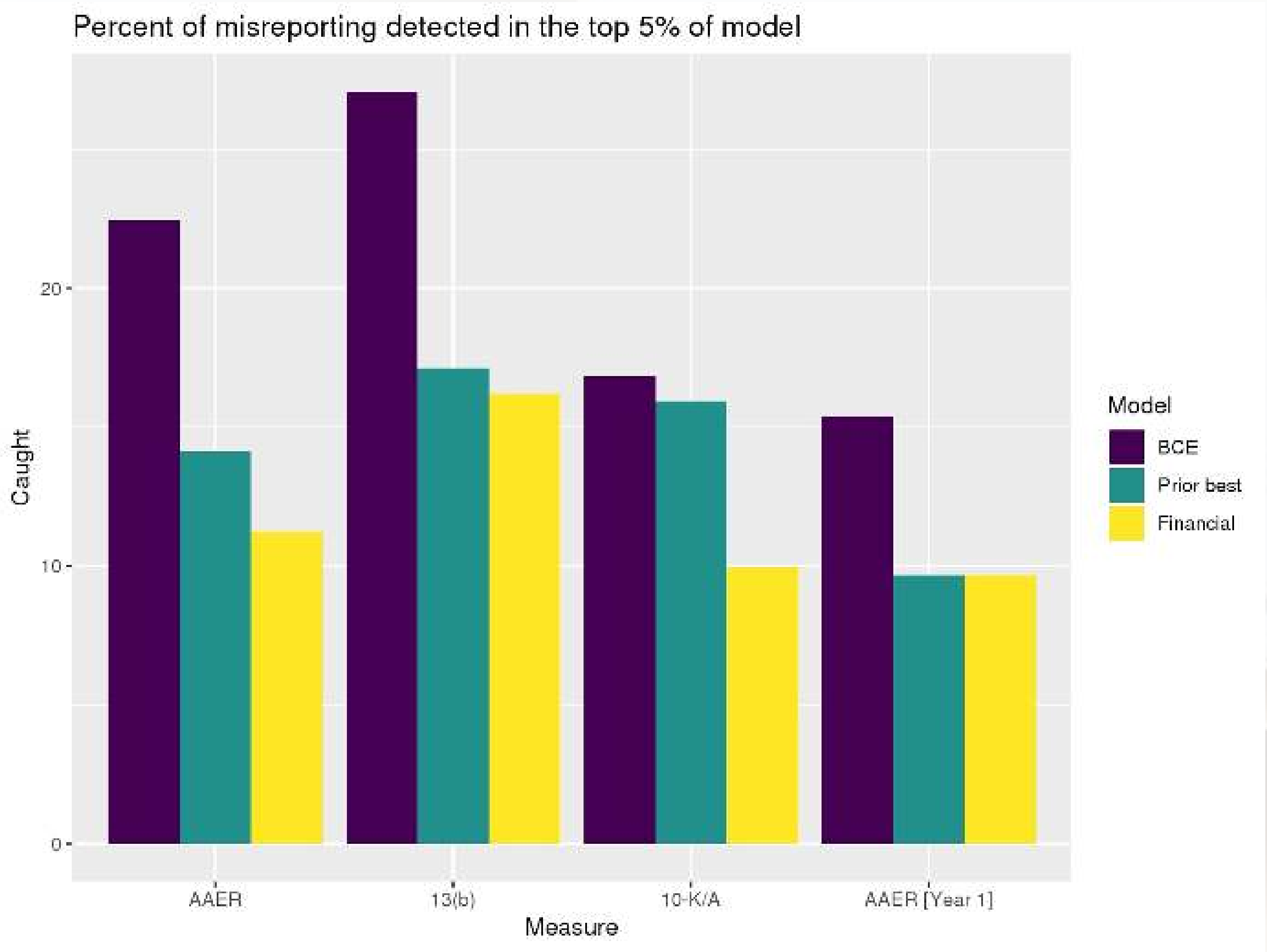
- 1990s: Financials and financial ratios
  - Misreporting firms' financials should be different than expected
- Late 2000s/early 2010s: Characteristics of firm disclosures
  - **Annual report** length, sentiment, word choice, …
- Late 2010s: More holistic text-based ML measures of disclosures
  - Modeling *what* the company discusses in their **annual report**

> All of these are discussed in Brown, Crowley and Elliott (2018) – I will refer to the paper as **BCE** for short

What we need to address:

1. Detecting varied events
   - "Careful" feature selection (offload to econometrics)
   - Intelligent feature design (partially offload to ML)
2. For business users… Interpretability matters
   - Psychology-style experiment
     - And a quasi-experiment
3. Predictive model
   - Need clean, out of sample designs + backtesting
   - Windowed design – data from 1998 won't help today, but it would in 1999
4. Infrequent events
   - Good for society, bad for modeling
   - Careful econometrics

Main results



Issue 1: Varied events

Past models

Financial model based on Dechow, et al. (2011)

- 17 measures including:
  - Log of assets
  - % change in cash sales
  - Indicator for mergers
- Theory: Purely economic
  - Misreporting firms' financials should be different than expected
    - Perhaps more income
    - Odd capital structure

Textual style model based on various papers

- 20 measures including:
  - Length and repetition
  - Sentiment
  - Grammar and structure
- Theory: Communications
  - Style reflects complexity and unintentional biases
  - Some measures ad hoc
  - Misreporting $\Rightarrow$ annual report written differently

> We tested an additional 26 financial & 60 style variables

The BCE model

1. Retain the variables from the previous models regressions
   - Forms a useful baseline
2. Add in an ML measure quantifying how much each **annual report** (~20-300 pages) talks about different *topics*
   - Train on windows of the prior 5 years
     - Balance data staleness, data availability, and quantity of text
     - Optimal to have 31 topics per 5 years
       - Based on in-sample logistic regression optimization

> Why do we do this? — Think like a fraudster!

- From communications and psychology:
  - When people are trying to deceive others, what they say is carefully picked – *topics* chosen are intentional
- Putting this in a business context:
  - If you are manipulating inventory, you don't talk about inventory

What the topics look like

Topic 6

Topic 11

Topic 21

Topic 30

Topic 2

Topic 9

Topic 12

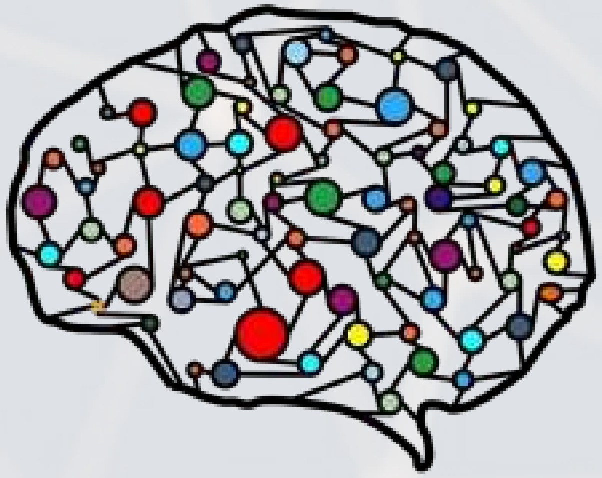
Topic 26

Topic 8

Topic 19

How to do this: LDA

- LDA: Latent Dirichlet Allocation
  - Widely-used in linguistics and information retrieval
    - Available in C, C++, Python, Mathematica, Java, R, Hadoop, Spark, …
    - We used onlineldavb
    - Gensim is great for python; STM is great for R
  - Used by Google and Bing to optimize internet searches
  - Used by Twitter and NYT for recommendations
- LDA reads documents all on its own! You just have to tell it how many topics to find



Implementation details

> The usual addage that data cleaning takes the longest still holds true

1. Annual reports are a mess
   - Fixed width text files; proper html; html exported from MS Word…
   - Embedded hex images
   - Solution: Regexes, regexes, regexes
     - Detailed in the paper’s web appendix
2. Stemming, tokenizing, stopwords
3. Feed to LDA
4. Tune hyperparameters (# of topics is most crucial)
5. Finally implement the model

Other considerations

1. LDA provides the *weight* on each topic, but documents vary a lot by length
   - Solution: Normalize to a percentage between 0 and 1
2. There is a mechanical component to topics due to firms' industries
   - Solution: Orthogonalize topics to industry
     - Run a linear regression and retain $\varepsilon_{i,firm}$:

$$topic_{i,firm} = \alpha + \sum_j \beta_{i,j} Industry_{j,firm} + \varepsilon_{i,firm}$$

Issue 2: Interpretability

LDA Verification

- LDA is well validated on general text, no question
- One key is to present some details of the topics to ensure comfort
- Another key is having prior evidence to fall back on
  - Whether LDA works on business-specific documents is not so well studied
    - Most studies just ask people whether they agree with the hand-coded topic categorizations

> We decided to fill this gap

Experimental design

Instrument: A word intrusion task

- Which word doesn't belong?

1. Commodity, Bank, Gold, Mining
2. Aircraft, Pharmaceutical, Drug, Manufacturing
3. Collateral, Iowa, Residential, Adjustable

Participants

- 100 individuals on Amazon Turk (20 questions each)
  - Human but not specialized

Quasi-experimental design

- 3 Computer algorithms (>10M questions each)
  - Not human but specialized
  1. GloVe on general website content
     - Less specific but more broad
  2. Word2vec trained on Wall Street Journal articles
     - More specific, business oriented
  3. Word2vec directly on annual reports
     - Most specific

> These learn the “meaning” of words in a given context

> Run the *exact same* experiment as on humans

Experimental results

Issue 3: Predictive modeling

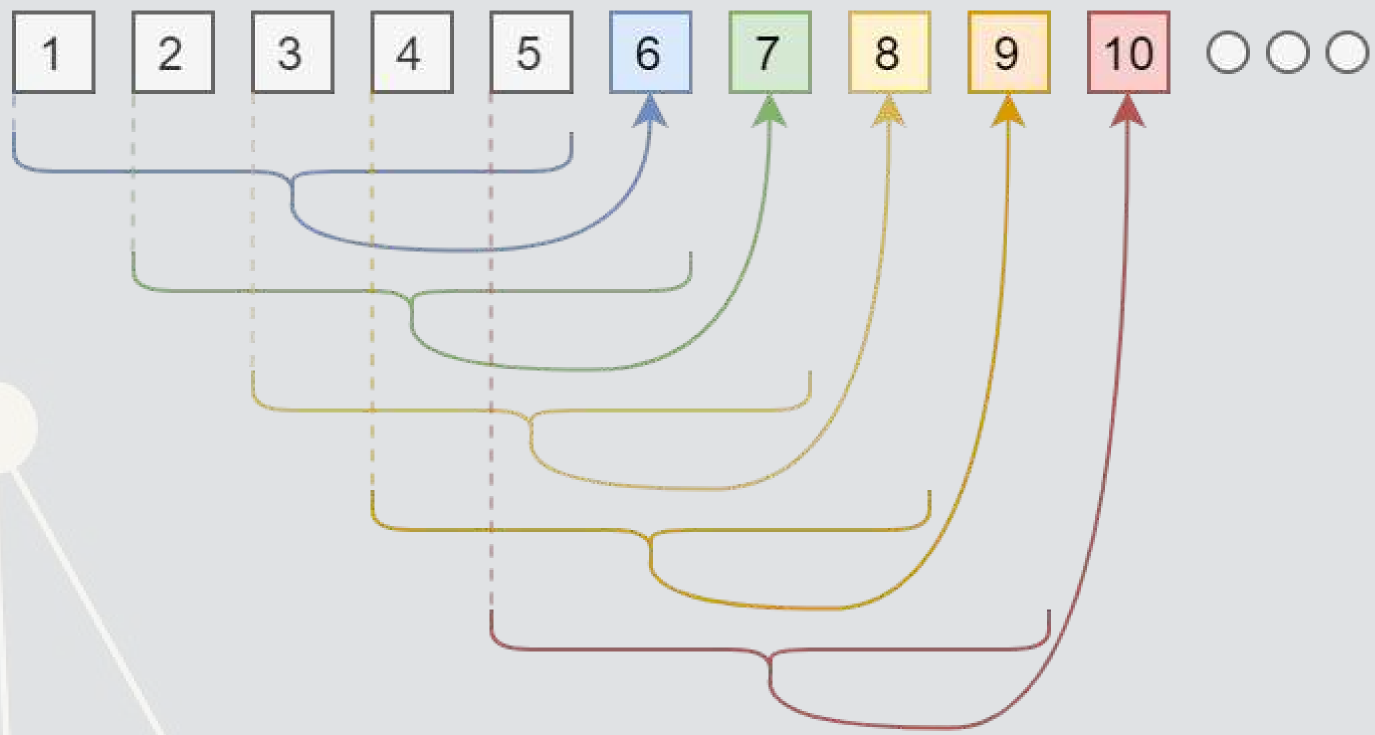
Backtesting

> We don't know who is misreporting today

- So, we will backtest
  - Use historical data to validate our model
- Problems:
  1. Misreporting changes over time
  2. Misreporting is unobservable (until it's observable)

Moving target

- Implement a moving window approach
  - 5 years for training + 1 year for testing
  - The study uses data from 1994 through 2012 – 14 possible windows
- Ex.: to predict misreporting in 2010, train on data from 2005 to 2009



> Problem: Now we have 14 models…

Comparing multiple models

- Performance measures:
  1. **ROC AUC**
  2. **Fisher statistics**
  3. Performance at a reasonable cutoff (5%)
  4. NDCG@k (usually used in ranking problems)

> ROC AUC and Fisher statistics will also allow us to statistically compare across models

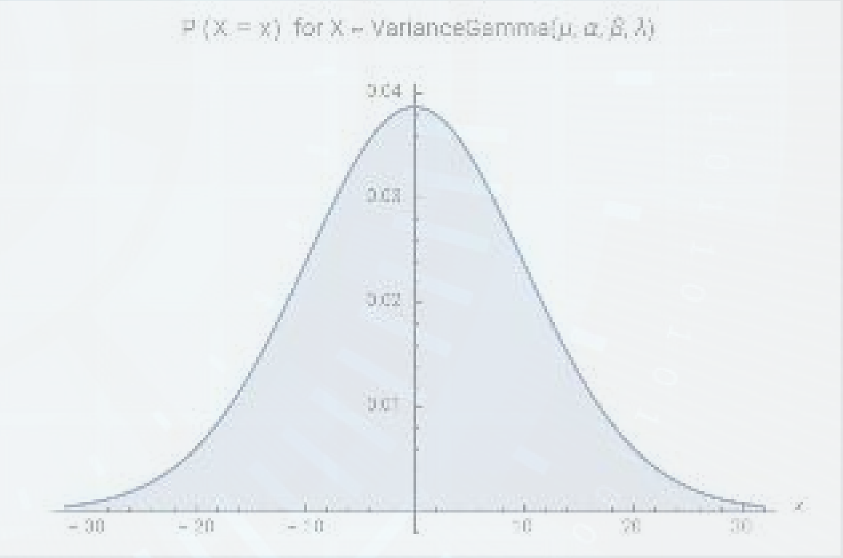
ROC AUC for windowed approaches

- ROC AUC
  - What is the probability that a randomly selected 1 is ranked higher than a randomly selected 0
- A good score is above 0.70
- Aggregating:
  - Simple: average AUC
  - More useful: Pool predictions together (with clustering by year)
- Comparing ROC AUCs
  - Not simple…
  - Wald statistic with bootstrapped variance estimates clustered by year
    - Implemented in Stata as `rocreg`

Purely statistical method

- Fisher statistic (Fisher 1932)
  - Combining p-values (Note: $p \sim U\left[0,1\right]$)
    - p-values come from our out-of-sample prediction model
  - Calculated as: $X = -2\sum_{i=1}^{k} ln(p_i)$

- Comparing models: Variance-Gamma test (see BCE)
  - Key insight: difference of X^2 vars has the same MGF as the Variance Gamma dist
- Calculation below
  - K is the modified Bessel function of the second kind



$$\mathbb{P}(X_1 > X_2) = \int_{-\infty}^{X_1 - X_2} \frac{1}{2^k \sqrt{\pi}\Gamma(k)} \left|z\right|^{k-\frac{1}{2}} K_{k-\frac{1}{2}}\left(\left|z\right|\right) dz$$

Observability

- The other issue is that, as of a given year, say 2009, we do not know every firm that was misreporting
  - We could build an algorithm with perfect information, but it may fall flat on current, noisy data!
  - It could also give us a false impression of an algorithm's effectiveness when backtesting
  - Misreporting can take a long time to discover: Zale's started in 2004, finished in 2009, and was disclosed in 2011!

> Solution: Censor our data to what was known at the point in time

- Use data on when a misreporting case was first disclosed
  - If the fraud wasn't known by the end of the window, train as if that was 0 (as it was unobservable back then)
  - Mimics our current situation

Issue 4: Infrequent events

Dealing with infrequent events

- Fraud is infrequent
  - E.g.: Out of 38,311 firm-years of data, there are 505 firm-years subject to AAERs
- Key issue: We may have more variables than events in a window…
  - Even if we don't, convergence is iffy using a logistic model
- A few ways to handle this:
  1. Very careful model selection (keep it sufficiently simple)
  2. Sophisticated degenerate variable identification criterion + simulation to implement complex models that are just barely simple enough
     - The main method in BCE
  3. Automated methodologies for pairing down models (LASSO, XGBoost)

Degenerate variable identification

1. Toss every input into a model
2. Check independentness using a QR decomposition
   - This will let us determine an order for dropping inputs
   - $A = Q \times R$, where A is our feature matrix, Q is an orthogonal matrix, and R is the transformation
     - More weight on the diagonal element in R means more independent (effectively)
     - Same underlying method as a Gram-Schmidt process
3. Remove excess inputs if too few 1s
   - Why? Because logit can't converge if there are more inputs than events (or non-events) in the data

> Independentness is a useful criterion for removing features with lower likelihood of being useful

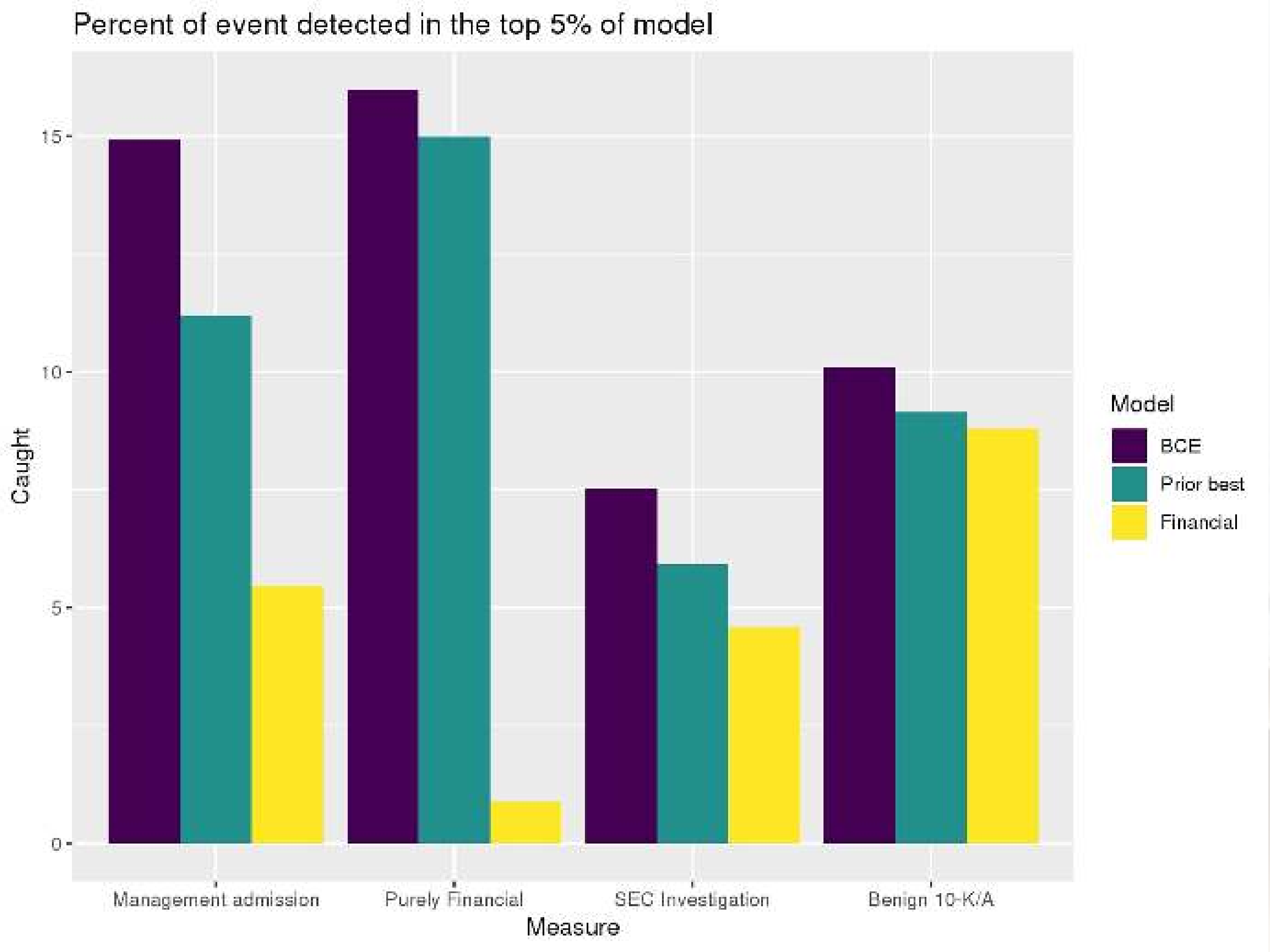
Logistic iteration

1. Run a logit using a Newton-Raphson solver for 50 iterations
2. Check convergence for signs of quasi-completeness
   - Standard errors will be in the millions if quasi-complete
   - If quasi-complete, drop the next least independent variable and restart
3. Run a 500 iteration logit using a Newton-Raphson solver
4. Recheck convergence
   - If failed, drop the next least independent variable and restart

> We will essentially get the most complex feasible model with the most independent set of features

Final comments

Some other interesting results



Ways to build on this model

1. Use a better tokenizer such as spaCy
   - Our tokenizer didn't detect noun phrases
2. Use econometric methods that are better suited for sparsity
   - E.g.: XGBoost
3. Consider using a more powerful LDA variant such as supervised LDA (sLDA)
4. No need to stop at LDA – there have been a lot of advancements in NLP since 2003

> Final note: The motivation behind our work was not to build a better mousetrap, but to illustrate the usefulness documents' content to better understand company/manager behavior

End matter

Thanks!

Dr. Richard M. Crowley
SMU

rcrowley@smu.edu.sg • @prof_rmc
Web: rmc.link

To learn more:

- These slides publicly available at rmc.link/DSSG
  - Plenty of links to click through and explore
- Technical details publicly available at SSRN

Case studies

- Prediction scores for **1999** ranked in the 98th percentile
  - First publicized in **2001**
- *Increases in Income* topic and firm size are the biggest red flags

- Prediction scores for **2004** through **2009** rank 97th percentile or higher each year
  - AAER published in **2011**
- *Media* and *Digital Services* topics are the red flags

Financial model

- Log of assets
- Total accruals
- % change in A/R
- % change in inventory
- % soft assets
- % change in sales from cash
- % change in ROA
- Indicator for stock/bond issuance
- Indicator for operating leases
- BV equity / MV equity
- Lag of stock return minus value weighted market return
- **Below are BCE's additions**
- Indicator for mergers
- Indicator for Big N auditor
- Indicator for medium size auditor
- Total financing raised
- Net amount of new capital raised
- Indicator for restructuring

> Based on Dechow, Ge, Larson and Sloan (2011)

Style model (late 2000s/early 2010s)

- Log of # of bullet points + 1
- # of characters in file header
- # of excess newlines
- Amount of html tags
- Length of cleaned file, characters
- Mean sentence length, words
- S.D. of word length
- S.D. of paragraph length (sentences)

- Word choice variation
- Readability
  - Coleman Liau Index
  - Fog Index
- % active voice sentences
- % passive voice sentences
- # of all cap words
- # of “!”
- # of “?”

> From a variety of research papers