PREDICTING ADVERSE DISCLOSURES IN CORPORATE FILINGS

ALEX NOCELLA

INTRODUCTION

- Goal
  - Make money
- Secondary goal
  - Algorithmically process text data in company filings
  - Detect anomalies
  - Use results to inform investment portfolio decisions

PRIOR WORK

- General process
  - Start with SEC Form 10-K
  - Identify fraud terms and features in the cross-section
  - Classify documents/companies as fraudulent
- Required data
  - Form 10-K
  - Tagged dataset of fraudulent documents
- List of things I did not have
  - Tagged dataset of fraudulent documents

PRIOR WORK (CONT'D)

- Typical accuracy around 80-90% in recent papers
  - Consider 50% baseline as many papers match class observations 1:1
- Common models used
  - SVM
  - Naïve Bayes
  - Logistic Regression
- Common feature techniques
  - TF-IDF token weights
  - Part of speech tagging
  - Word counts

PRIOR WORK (CONT'D)

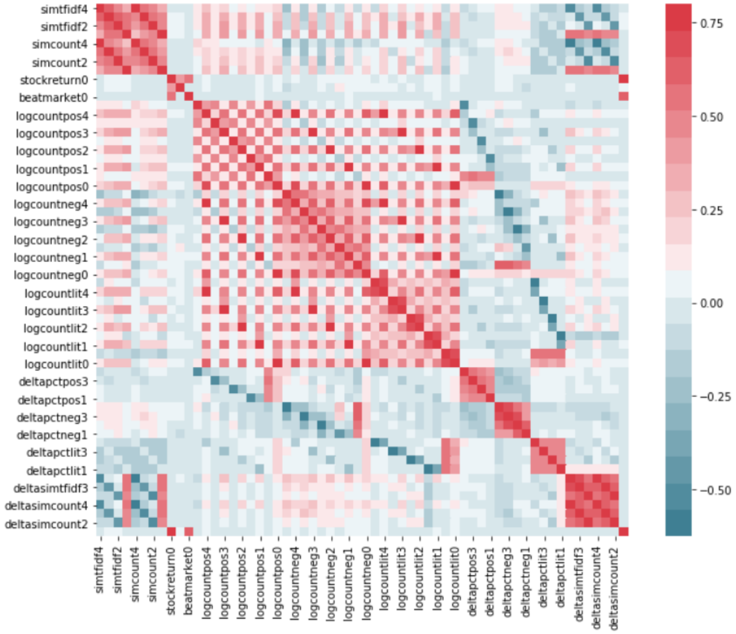
- Li (2006) has a similar idea
  - Go through all the Form 10-K filings you can find
  - Count the number of "risk" or "uncertainty"-like words, apply log
    - Literally, the paper just mentions things like "risk," "risks," "risky," "uncertain", "uncertainty"
  - Account for a lot of other company attributes
    - I don't do this step
    - Fama-French + momentum model
    - Throw away financial companies because seeing "risk" in Form 10-K is not meaningful
  - Paper claims 10% annualized alpha (excess return)
    - Buy low-change-in-risk-log-count names, sell high
    - Big if true

MY WORK

- SEC Form 10-Q and 10-K for current S&P 500 constituents
- Go back as far as possible (earliest 1994)
- Get rid of everything that isn't a word (100K+ word dictionary)
- Get rid of stop words
- Vectorize: TF-IDF and Count
- Count positive/negative/litigious tagged words (couple thousand word dictionary)
- Compute cosine similarity for both vectorizations, today vs each of prior 4 filings
- Compute changes in cosine similarities, tagged word counts, tagged word frequencies
- Target is whether stock outperforms market between now and next file date

SIZE OF DATA

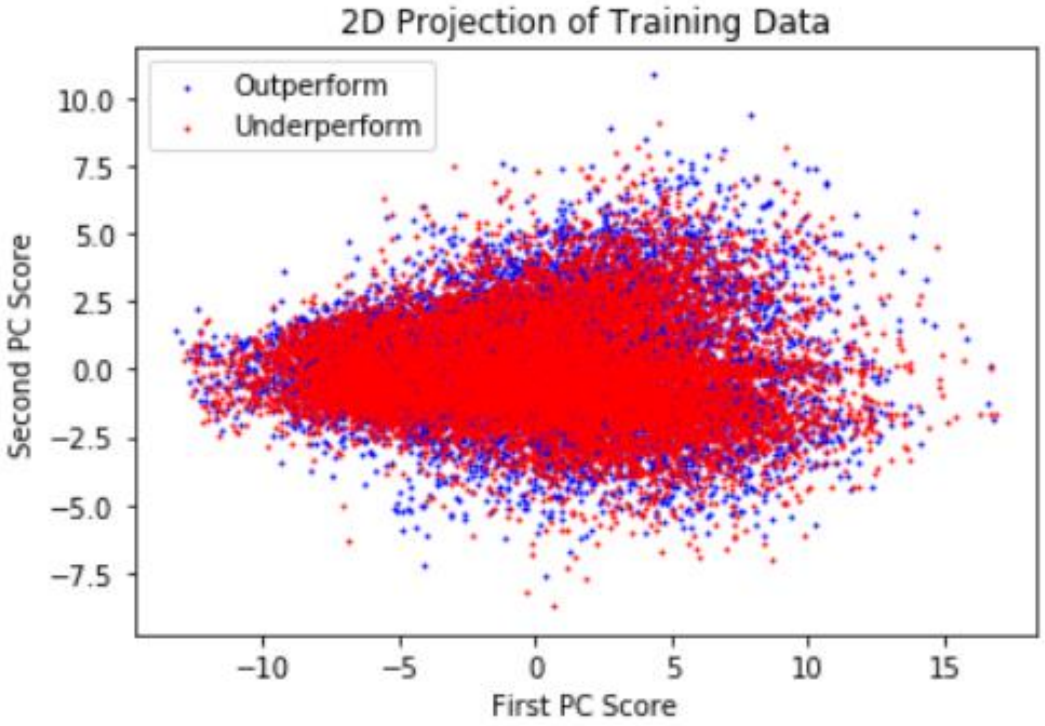
- Roughly 38,000 rows (row = 1 form filing, aka firm-quarter)
- 60 features



MACHINE LEARNING

- Binomial response variable
- Test data
  - Filing date 2015-present
  - > 5,000 records
- Dev data
  - Randomized 30% of the ~33,000 remaining records
- Easy candidates for linearly separable data
  - Logistic regression
  - SVM
  - PCA

LINEARLY SEPARABLE?



NONLINEAR ATTEMPTS

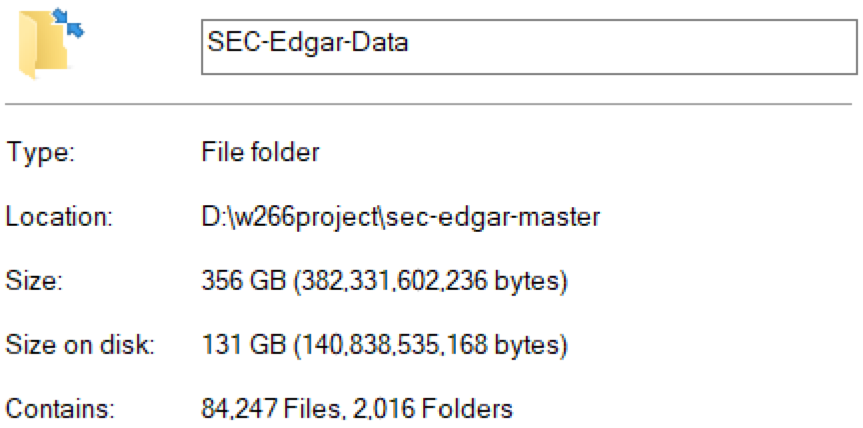
- K-nearest neighbors
  - Frequently worse than guessing one class every time
- Naïve Bayes
  - In most trials, identical accuracy to guessing one class every time
- Decision trees
  - Had trouble generalizing – extremely impressive, but overfit, accuracy

NONLINEAR ATTEMPTS (CONT'D)

- Multilayer perceptron
  - 4 hidden layers each with 100 neurons
  - Still suffered from overfitting
  - Dev data
    - 52.7% accuracy conditional / 53.1% unconditional
    - 57.0% precision conditional / 50.1% unconditional
    - 57.7% recall conditional / 39.3% unconditional
  - Test data
    - 52.3% accuracy conditional / 50.4% conditional
    - 56.7% precision conditional / 54.2% conditional
    - 17.7% recall conditional / 14.1% unconditional
    - Only 671 predictions out of > 3,000 samples in conditioned set (predicted prob > 0.64)

NOT SO FUN FACTS

- Akamai has their own binomial classification system
  - The amount of data I downloaded from the SEC led to an IP ban
  - I learned that Akamai hosts tons of content, like comcast.com, fidelity.com, vue.playstation.com



SEC-Edgar-Data

Type:	File folder
Location:	D:\w266project\sec-edgar-master
Size:	356 GB (382,331,602,236 bytes)
Size on disk:	131 GB (140,838,535,168 bytes)
Contains:	84,247 Files, 2,016 Folders

REFERENCES

- Sean L. Humpherys, Kevin C. Moffitt, Mary B. Burns, Judee K. Burgoon, William F. Felix, Identification of fraudulent financial statements using linguistic credibility analysis, In Decision Support Systems, Volume 50, Issue 3, 2011, Pages 585-594, ISSN 0167-9236, https://doi.org/10.1016/j.dss.2010.08.009
- Yuh-Jen Chen, Chun-Han Wu, Yuh-Min Chen, Hsin-Ying Li, Huei-Kuen Chen, Enhancement of fraud detection for narratives in annual reports, In International Journal of Accounting Information Systems, Volume 26, 2017, Pages 32-45, ISSN 1467-0895, https://doi.org/10.1016/j.accinf.2017.06.004
- Sunita Goel, Jagdish Gangolly, Sue R. Faerman, and Ozlem Uzuner (2010) Can Linguistic Predictors Detect Fraudulent Financial Filings?. Journal of Emerging Technologies in Accounting: December 2010, Vol. 7, No. 1, pp. 25-46. https://doi.org/10.2308/jeta.2010.7.1.25
- Li, Feng, Do Stock Market Investors Understand the Risk Sentiment of Corporate Annual Reports? (April 21, 2006). http://dx.doi.org/10.2139/ssrn.898181