Discussion of "Big Data Analyses with No Digital Footprints Available – Evidence from Cyber-Telecom Fraud"

by Liu, Liu, Ruan, Yang, and Zhang

CICF 2021

Keer Yang – University of Minnesota

Summary and Main Contributions

- This paper studies cyber-telecom fraud and the effectiveness of big data and machine learning techniques in identifying these cyber-telecom fraud.
  - Female borrowers are more likely to be fraud victims.
  - Big data and ML algorithms increase fraud detection accuracy, even when no digital footprints available.

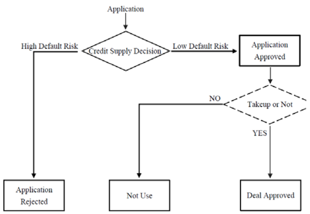
- Very important research question!

- Contributions:
  - Fraud in the FinTech era
    - FinTech brings in efficiency
    - Fraud impedes borrowers' use of FinTech
    - Important to understand who are more likely to be fraud victims and how to prevent cyber-telecom fraud
  - Big Data and ML in Finance
    - Increasing in the predictive power of ML + Big Data improves the efficiency of the market

My discussion

- Comments 1 : understand cyber-telecom fraud
- Comments 2 and 3: understand the role of ML algorithm and big data

Comment 1: Commission of Fraud vs Reporting of Fraud



Panel A: Male VS Female

	Male, Control	Female, Control
N	126,847	60,332
No. of Fraud	46	269
No. of Use	44	257
Prob. of Fraud	0.0363%	0.4459%
Prob. of Use conditional on Fraud	95.65%	95.54%

- Fraud is self-reported
  - post-borrowing feedback (in treated and control groups)
  - feedback from warning calls (in treated group)
- The authors find that female borrowers are more likely to be fraud victim.
- Or female borrowers are more likely to report fraud?

Comment 1: Commission vs Reporting

- P(Observed Fraud) = P(Commission of Fraud)*P(Reporting Fraud) (Wang, Winton, and Yu (2010),Wang (2013))

Case One:

- Female borrowers are more likely to report fraud

	Female	Male
Commission of Fraud	100	100
Reporting of Fraud	80	20

Case Two:

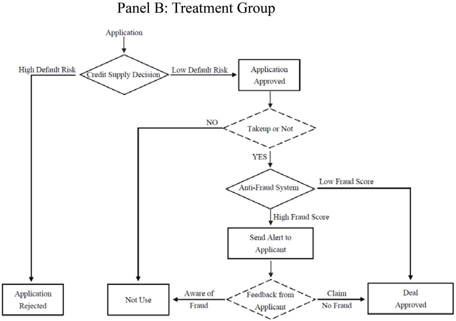
- If male are rejected more by credit decision
- In rejected loans, 0% report fraud
  In approved loans, 10% report fraud

Reporting of Fraud	Female		Male	
	No	Yes	No	Yes
Rejected	10	0	50	0
Approved	81	9	45	5
Total	9		5	

Comment 1: Commission vs Reporting

- Solution for different report probability among rejected and approved loans
  - compare fraud rates between approved and rejected loans
  - compare loan rejection rates across different groups
- Solution for different report probability among female and male borrowers
  - survey?

Comment 2: How do ML + Big Data help?



- Anti-fraud system: (1) use ML + Big Data to select; (2) make warning calls
- Anti-fraud system has a model accuracy of 2.6%
  - larger than 0.18%, which is population probability of observing fraud
- Warning calls increase fraud reporting?
- ML+ Big Data select actual fraud, or reported fraud (interesting to know)

Comment 2: How do ML + Big Data help?

- Population and No Anti-fraud system
- No improvement in detecting fraud commission
- Improvement in detecting fraud commission, only concentrated in reported fraud
- Improvement in detecting fraud commission

	Population	Random	ML + Big Data		
		Drawing	Case 1	Case 2	Case 3
Reported Fraud	50	5	5	8	10
Unreported Fraud	50	5	5	2	10
Not Fraud	900	90	90	90	80
Total Number	1000	100	100	100	100
No Warning Calls — Unreported Fraud won't be identified					
Model Accuracy		5%	5%	8%	10%
Warning Calls — Unreported Fraud will be identified					
Model Accuracy		10%	10%	10%	20%

Comment 2: How do ML + Big Data help?

- It is possible that ML + Big data do not improve fraud detection rate (case 1), or just select fraud-induced loans that are more likely to be reported
  - If so, simply random calls can achieve the same better performing
- Evidence from the back-test results, model accuracy is slightly lower than the treatment group (1.59% vs 2.60%)
  - Rule out "no improvement" (case 1)
  - 2.60% is much larger than 0.18% (sample average fraud rate), partially rule out the possibility of "only selecting fraud-induced loans that are more likely to be reported" (case 2)
- One possible solution to case 2: make similar warning calls in a randomly selected group, see if there is a difference in model accuracy

Comment 3: ML + Heterogeneity = Distributional Consequences ?

- If we use observed fraud to train the ML model, the model may only benefit borrowers who are more likely to report the loans
- Borrowers who are more likely to report will be selected and warned by the anti-fraud system, whereas borrowers who do not report will not benefit from the anti-fraud system
- Distributional consequences?
- Distributional consequences of better statistical technology have been documented in credit decisions (Fuster et al. (Forthcoming))

Conclusion

- Fascinating Paper!
- Help us understand cyber-telecom fraud and the role of ML and big data.
- Hope my comments will help with the next version of the paper.

Appendix

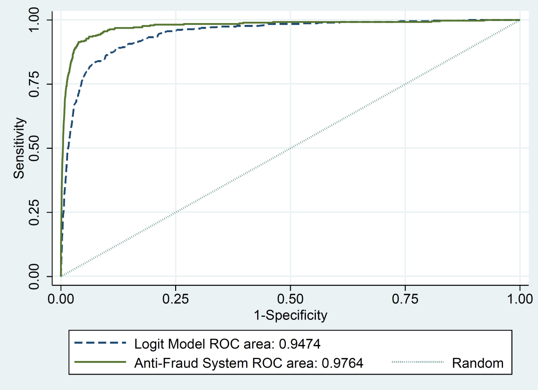
References I

Fuster, A., Goldsmith-Pinkham, P., Ramadorai, T., Walther, A., Forthcoming. Predictably unequal? the effects of machine learning on credit markets. Journal of Finance .

Wang, T. Y., 2013. Corporate securities fraud: Insights from a new empirical framework. The Journal of Law, Economics, & Organization 29, 535–568.

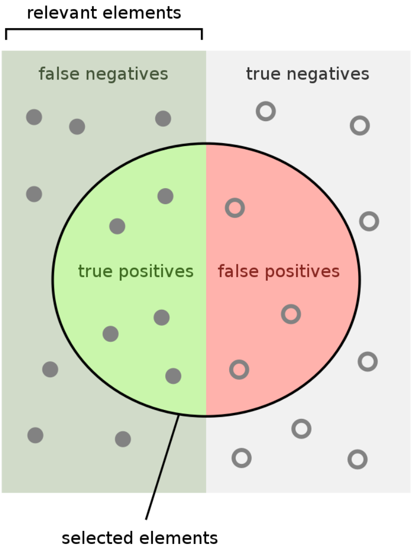
Wang, T. Y., Winton, A., Yu, X., 2010. Corporate fraud and business conditions: Evidence from ipos. The Journal of Finance 65, 2255–2292.

Comment: How much does ML +Big Data h?



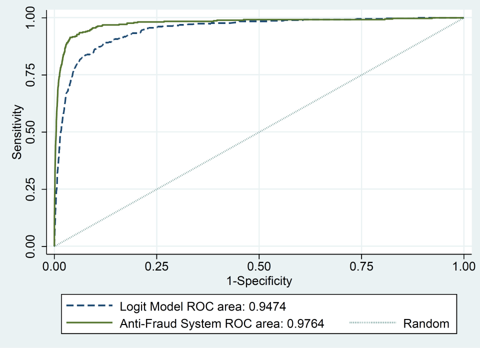
- Comparing treated group (with anti-fraud system) to OLS anti-fraud algorithm
- Plot the ROC curve
  - Sensitivity: True Positives
  - Specificity : True Negatives
- and calculate the area under the curve

Comment 3: How much does ML +Big Data h?



- My comment focus on the second comparison.
- What determines the effectiveness of a anti-fraud system, assume total number N, fraud rate = f
  - model accuracy + detection rate
- What determines the goodness of a predictive model?
  - Sensitivity + Specification
- Sensitivity + Specification + positive rate (fraud rate, f) determines model accuracy + detection rate

Comment : How much does ML +Big Data h?



- Cost and Benefit Calculation, assume total number N, fraud rate = f
- Benefit: Number of case correctly identified
  = detection*N*f
  = sensivity*N*f
- Cost: Number of identified
  = detection*N*f/accuracy
  = sensivity*N*f + N*f*(1 - specification)*(1-f)/f
- Moreover, specification does not affect total benefits
- Need a weighted version of "AUC"

Comment : How much does ML +Big Data help?

- Anti-fraud system: GRBT + Big Data
- Back Testing in Figure 4 (also in Figure 5?): OLS + Small Data

(accuracy rate, detection rate)

	Small Data	Big Data
OLS	(1.59%,89.21%)	?
GRBT	?	(2.60%,89.87%)

- Similar detection rate, higher accuracy
- How much improvement from big data?
- How much improvement from GRBT?