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

- 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

- $P(\text{Observed Fraud}) = P(\text{Commission of Fraud}) \times P(\text{Reporting Fraud})$
  
  (Wang, Winton, and Yu (2010), Wang (2013))

Case One:
- Female borrowers are more likely to report fraud

<table>
<thead>
<tr>
<th></th>
<th>Female</th>
<th>Male</th>
</tr>
</thead>
<tbody>
<tr>
<td>Commission of Fraud</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Reporting of Fraud</td>
<td>80</td>
<td>20</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th></th>
<th>Female</th>
<th>Male</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reporting of Fraud</td>
<td>No Yes</td>
<td>No Yes</td>
</tr>
<tr>
<td>Rejected</td>
<td>10 0</td>
<td>50 0</td>
</tr>
<tr>
<td>Approved</td>
<td>81 9</td>
<td>45 5</td>
</tr>
<tr>
<td>Total</td>
<td>9 5</td>
<td></td>
</tr>
</tbody>
</table>
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

<table>
<thead>
<tr>
<th></th>
<th>Population</th>
<th>Random Drawing</th>
<th>ML + Big Data</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Case 1</td>
<td>Case 2</td>
</tr>
<tr>
<td>Reported Fraud</td>
<td>50</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>Unreported Fraud</td>
<td>50</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>Not Fraud</td>
<td>900</td>
<td>90</td>
<td>90</td>
</tr>
<tr>
<td>Total Number</td>
<td>1000</td>
<td>100</td>
<td>100</td>
</tr>
</tbody>
</table>

No Warning Calls — Unreported Fraud won’t be identified

Model Accuracy

<table>
<thead>
<tr>
<th></th>
<th>Population</th>
<th>Random Drawing</th>
<th>ML + Big Data</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>5%</td>
<td>5%</td>
<td>8%</td>
</tr>
</tbody>
</table>

Warning Calls — Unreported Fraud will be identified

Model Accuracy

<table>
<thead>
<tr>
<th></th>
<th>Population</th>
<th>Random Drawing</th>
<th>ML + Big Data</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>10%</td>
<td>10%</td>
<td>10%</td>
</tr>
</tbody>
</table>
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


Comment: How much does ML + Big Data help?

- 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 help?

- My comment focuses on the second comparison.

- What determines the effectiveness of an anti-fraud system, assuming a 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 help?

- **Cost and Benefit Calculation**, assume total number $N$, fraud rate $= f$

- **Benefit**: Number of case correctly identified
  
  $= \text{detection} \times N \times f$
  
  $= \text{sensivity} \times N \times f$

- **Cost**: Number of identified
  
  $= \text{detection} \times N \times f / \text{accuracy}$
  
  $= \text{sensivity} \times N \times f + N \times f \times (1 - \text{specification}) \times (1 - f) / f$

- 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

<table>
<thead>
<tr>
<th>(accuracy rate, detection rate)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Small Data</td>
</tr>
<tr>
<td>OLS</td>
</tr>
<tr>
<td>GRBT</td>
</tr>
</tbody>
</table>

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