I KNOW COMPANY
WHAT YOUR DID
LAST SUMMER

CORPORATE RESIDENCE FRAUD DETECTION
USING FINE-GRANED PAYMENT DATA

Bart Minnaert
Marija Stankova
David Martens
Julie Moeyersoms
Enric Junqué de Fortuny
Foster Provost

3:30PM-5PM | NY Ballroom East
Problem setting

Very high corporate tax rates in Belgium \[^1\]

Avoiding Belgian tax by falsely putting the head office in a country with lower corporate taxes.

= Corporate residence fraud

Problem setting: relevance?

“Fraud is as Belgian as beer and fries”

Belgian State Secretary for Fraud

European Commission estimates:

- Loss of €30 billion annually due to fiscal fraud in Belgium [2]
- Overall European losses up to €1 trillion [3]

Problem setting

Fairway Golf club
Raw data overview

**Structured data**

Examples: sector, address, ...

→ For Belgian companies (Foreign companies: only country)

**Transactional data**

Invoices between foreign and Belgian companies

- 6,802,710 transactions
- 108,753 Belgian companies
- 858,703 Foreign companies
- 18.66 avg / Belgian company
- 2.36 avg / Foreign company
Overview of the predictive system

Structured Data + Transactional Data

Structured Data

- Scalable methods: SVM, NV, SVM

Transaction Models

- Probability Scores

Stacked Model

classification methods: SVM
Structured data

Combine structured data and high level data:

**Structured data**: company characteristics
   a) Foreign Location: *LUX*
   b) Target Label: *Yes/No/?*

**Extracted structured data**: from transactional data
   a) Most frequent location: *Charleroi*
   b) Transaction value (avg): €79,053.5
   c) ...
Transactional data: fine-grained matrix

\[
\begin{bmatrix}
1 & 0 & 0 & 0 \\
0 & 1 & 1 & 0 \\
1 & 0 & 1 & 1 \\
0 & 0 & 0 & 1 \\
\end{bmatrix}
\rightarrow
\begin{bmatrix}
1 \\
0 \\
0 \\
1 \\
\end{bmatrix}
\]
Transactional data: wvRN

Transactional data: wvRN

Predictive power

<table>
<thead>
<tr>
<th>Technique used</th>
<th>Combined data</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AUC</td>
<td>std.dev.</td>
<td></td>
</tr>
<tr>
<td>wvRN$_T$</td>
<td>94.55</td>
<td>(±5.26)</td>
<td></td>
</tr>
<tr>
<td>Naive Bayes$_T$</td>
<td>94.74</td>
<td>(±5.45)</td>
<td></td>
</tr>
<tr>
<td>SVM$_T$</td>
<td>70.26</td>
<td>(±12.46)</td>
<td></td>
</tr>
<tr>
<td>SVM$_T$ (50-50)</td>
<td>74.85</td>
<td>(±19.97)</td>
<td></td>
</tr>
<tr>
<td>SVM$_S$</td>
<td>91.77</td>
<td>(±8.16)</td>
<td></td>
</tr>
<tr>
<td>SVM$_S+T$</td>
<td>96.22</td>
<td>(±4.8)</td>
<td></td>
</tr>
</tbody>
</table>
Predictive power

- Up to 100% improvement in detection rate
Comprehensibility: a problem
Instance based explanations

IF

ACME Shady Products

DID NOT

receive an invoice from the Fairway Golf Club

THEN

the predicted class would change to non-fraudulent

✓ Identify the most significant reason for prediction

Instance based explanations

IF

Rolex

DID NOT receive an invoice from the Fairway Golf Club

THEN

the predicted class would change to non-fraudulent

✓ Identify the most significant reason for prediction
✓ Allows for easy overruling
✓ Avoid disillusionment with the system and reluctance to use it when it makes mistakes

Conclusion

- Belgium is a fun country
- Fine-grained data potentially holds important information
- Explaining models for Big Data: instance based
“The interaction between the two worlds [academia and government] has proven very valuable. Other countries are now visiting Belgium to see how the Social Intelligence and Investigation Service and the Special Tax Inspection service apply this technique. That is why we need to continue to invest in this technology.”

Belgian State Secretary for Fraud