Machine Learning for Fraud Detection

Dídac Rodríguez Arbonès
didac@di.ku.dk
Fraud

- Fraud in Denmark is a substantial issue.
- The financial institutions are keen on prevention.
- It is a very difficult problem to solve.
- Substantial resources are dedicated to fight against it.

<table>
<thead>
<tr>
<th>Year</th>
<th>1. kvartal</th>
<th>2. kvartal</th>
<th>3. kvartal</th>
<th>4. kvartal</th>
<th>I alt</th>
</tr>
</thead>
<tbody>
<tr>
<td>2013</td>
<td>Netbankindbrud</td>
<td>35</td>
<td>60</td>
<td>40</td>
<td>34</td>
</tr>
<tr>
<td></td>
<td>- Heraf med tab</td>
<td>14</td>
<td>26</td>
<td>14</td>
<td>17</td>
</tr>
<tr>
<td></td>
<td>Tabets størrelse i kroner</td>
<td>607,366,20</td>
<td>1,601,254,11</td>
<td>811,590,14</td>
<td>2,288,305,83</td>
</tr>
<tr>
<td>2012</td>
<td>Netbankindbrud</td>
<td>29</td>
<td>66</td>
<td>58</td>
<td>46</td>
</tr>
<tr>
<td></td>
<td>- Heraf med tab</td>
<td>10</td>
<td>12</td>
<td>24</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>Tabets størrelse i kroner</td>
<td>2,316,543</td>
<td>375,213,99</td>
<td>1,460,753</td>
<td>2,157,237</td>
</tr>
<tr>
<td>2011</td>
<td>Netbankindbrud</td>
<td>0</td>
<td>0</td>
<td>8</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>- Heraf med tab</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Tabets størrelse i kroner</td>
<td>0</td>
<td>0</td>
<td>23,400</td>
<td>136,268</td>
</tr>
<tr>
<td>2010</td>
<td>Netbankindbrud</td>
<td>0</td>
<td>0</td>
<td>12</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>- Heraf med tab</td>
<td>0</td>
<td>0</td>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Tabets størrelse i kroner</td>
<td>0</td>
<td>0</td>
<td>433,043</td>
<td>0</td>
</tr>
</tbody>
</table>

http://www.finansraadet.dk/tal--fakta/Pages/statistik-og-tal/netbankindbrud---statistik.aspx
Coverage

bankID
6 M users
~500 M transactions

NEM ID
4.5 M users
~145.5 M transactions
Overview of project

- Develop a prevention mechanism against online banking cyber-crime using machine learning (ML).
- Detection capable of identifying fraudulent authentication sessions in NemID.
- Possibility of working within other company IT assets.
- Project to be deployed to BankID in Norway and NemID in Denmark adding to the security measures currently in place.
- Self-updating and minimal operation intervention.
Phishing

It consists on redirecting the user to a malicious webpage.

The user is unaware who is the information recipient.

The criminals can use the credentials to authenticate to the systems.
Machine learning

• Part of artificial intelligence.
• Ecompasses many algorithms to learn from data and make decisions on new events.
• Its power comes from the available information from the system.
• There is no one-size-fits-all, it is as good as the data it relies on.
• There are two major types:
  - Supervised
  - Unsupervised
Machine learning

- Various levels of performance and requirements.
- Ensemble systems possible.
- Feature space reduction.
Data source

- The data is preprocessed and anonymized after it is collected from the login applet of the BankID.

- The data is complex and heterogenous:
  - Geographic information
  - Browser information
  - OS information
  - Timing information

- There are no reliable labels to be used for analysis.
Data source

- An example of user data are keystrokes:

  ![Diagram of keystrokes with timestamps](image)

- There have been previous studies achieving good results in classification only with this type of data.

- It relies on heavily on previous user knowledge.

- Unfortunately, this type of data is not enough on its own for a reliable fraud detection.

Data source

- Real-valued vectors
- Discrete data
- Unmatched sets
- Device information
- Usage data
System overview
Other information sources

- Credit cards
- Individual transactions
  - Person-to-person
  - Crowd-to-person
- Web service usage
- Customer profiling
Challenges & expectations

- Thieves are smart.
- Sometimes *very* smart.
- There exist numerous companies around the globe with the only goal of illicit financial gain.
- It is *humanly* impossible to detect and stop them without unlimited resources.
- Legal & marketing concerns make datasets scarce and incomplete.

- Machine learning is an excellent candidate to pursue these actions in a scalable manner with low expense and effort.
- It can find patterns not directly obvious to humans.
- It can self-adapt to new events.
Industrial PhD practicalities

- The programme is funded by the Ministry of Higher education and Science (DASTI). 
- There are several application reviews during the year. It typically takes 2-4 months to receive an answer.
- The company is required to have physical presence in Denmark.
- The student is formally employed at the company.
- There is a supervisor both at the university and at the company.
- Time allocation is roughly 50%, if not otherwise agreed.
- There is a budget allocated for the university expenses and a subvention for the company.
- Usually any patent or profitable outcome from the project is property of the company (IP agreement), but there has to be a scientific dissemination component.

(1) http://ufm.dk/en