Unsupervised Anomaly Detection
Claudi Ruiz Camps

Machine Learning Specialist
claudi.ruiz.camps@nl.abnamro.com
Business Goal

Detect anomalies at their earliest stage and avoid bigger problems in advance.
An anomaly is a data point which is significantly different from the remaining data.
Identifying Anomalies

Given a Raw Dataset that:

- Increases around 10 million data-points every day
- Changes its patterns over time
- Contains a mixture of text, categorical and numerical data types
Artificial Intelligence

Machine Learning

Deep Learning
Why Deep Learning

Deep learning models can

- Learn patterns in significantly large amounts of data
- Enable you to solve problems even if you are not an expert in the domain
- Approximate (theoretically) any function
Supervised vs Unsupervised Learning

**Supervised**

We do have labels

**Unsupervised**

We do not have labels (in our case we do not have explicit examples of anomalies)
Deep Learning

input layer  hidden layer 1  hidden layer 2  hidden layer 3

output layer
Autoencoders

- **Input**
  - Sample
- **Hidden Layers**
  - Code
- **Output**
  - Reconstructed Sample

- **Encoder**
- **Decoder**
Autoencoders

Input \( Y \) → Hidden Layers \( z \) → Output \( \hat{Y} \)

Sample \( Y \)

Encoder

Decoder

MSE

Reconstructed Sample
Probabilistic Autoencoder

The Autoencoder scores each sample based on its reconstruction error:

\[ \text{↑ reconstruction error} \implies \text{↑ probability to be an anomaly} \]
\[ \text{↓ reconstruction error} \implies \text{↓ probability to be an anomaly} \]

The Autoencoder won’t be able to properly reconstruct abnormal samples because they deviate so much from the other samples, they are generated by a different mechanism.
## Scores

<table>
<thead>
<tr>
<th>Sample id</th>
<th>Rec_Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>313971</td>
<td>1020</td>
</tr>
<tr>
<td>72302</td>
<td>980</td>
</tr>
<tr>
<td>115924</td>
<td>965</td>
</tr>
<tr>
<td>94811</td>
<td>853</td>
</tr>
<tr>
<td>236924</td>
<td>744</td>
</tr>
<tr>
<td>310405</td>
<td>729</td>
</tr>
<tr>
<td>172925</td>
<td>721</td>
</tr>
<tr>
<td>172920</td>
<td>666</td>
</tr>
<tr>
<td>236938</td>
<td>656</td>
</tr>
<tr>
<td>11831</td>
<td>324</td>
</tr>
<tr>
<td>313982</td>
<td>298</td>
</tr>
<tr>
<td>69467</td>
<td>265</td>
</tr>
<tr>
<td>69472</td>
<td>228</td>
</tr>
<tr>
<td>313974</td>
<td>219</td>
</tr>
<tr>
<td>313969</td>
<td>214</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Example
Example
Example

Input

Output
Example

Input

Output
Example

Input

Output

Anomaly!

High MSE
Software

Python
Main Programming Language

Spark
Fast and general engine for large-scale data processing

Nvidia
Parallel computing platform

TensorFlow
Open-source Machine Learning library from Google to build Deep Learning models
Our Own Application Programming Interface

Building our own Machine Learning API

- Directly connected to the production environment
- Easy to use for non-experts in machine learning in their own projects
- Always updated to the state of the art
Hardware: Deep Learning on a GPU machine

**CPU**
- Intel® Core™ i7-7700K processor

**GPUs**
- 2 x GeForce GTX 1080

**RAM Memory**
- 32 GB 2.667 MHz
Hardware: Deep Learning on a GPU machine (very soon)

CPU
2 x Intel Xeon Platinum 8168

GPUs
3 x V100

RAM Memory
512 GB 2.667 MHz
Hardware: Data Pre-processing on Spark

4 x Workers:
40 cores each
216 GB each

Master:
40 cores
216 GB
Thank you for your Attention!