Providing assurance with “Data Driven Assurance”

Audience: KNIME Talks 2021
Who am I?

Thanks for having me...

Michiel Krol
Head of Audit Data Excellence bij Rabobank Groep
Why started with DDA and what are the benefits

Digital world drives need for NextGen Audit

The digital world is changing rapidly and is becoming more and more data driven

This also has a profound impact on the Audit profession, on Audit services and on auditors

Audit Rabobank has set three ambitious data driven targets

-10/20% HOURS

ASSURANCE LEVELS UP IN TWO IN THREE AUDITS

DATASETS READILY AVAILABLE FOR Q&A
DDA is 90% smart business intelligence and 10% advanced analytics applied in the audit cycle.

How does DDA work

How often? How much?
Descriptives & Dashboarding

Is this complete and correct?
Data quality profiling & Reconciliation

How does this compare?
Benchmarking

Is this logical?
Business rules & Smart Stitching

How does this change?
Trend analysis

What steps, who, how long?
Process mining

Is this weird?
Anomaly detection

What do the words reveal?
Text analysis

How is this organized? (1)
Clustering analysis

How is this organized? (2)
Decision tree

How much? How many?
Regression analysis

Is this A or B?
Classification

21°C
It is a Target Operating Model rather than only a team of specialists

How does DDA work

<table>
<thead>
<tr>
<th>ORGANISATION</th>
<th>PEOPLE</th>
<th>DATA</th>
<th>TECHNOLOGY</th>
</tr>
</thead>
<tbody>
<tr>
<td>Operating Model</td>
<td>Staffing</td>
<td>Data Acquisition</td>
<td>Architecture (e.g. Data Lake)</td>
</tr>
<tr>
<td>Innovation Cycle</td>
<td>Learning &amp; Development</td>
<td>Data Governance &amp; Operations</td>
<td>Tooling (e.g. analysis, visualization)</td>
</tr>
<tr>
<td>Ways of working</td>
<td>Data Driven Culture</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
**How does DDA work**

**Start small... scale fast**

- **Start small**
  - Test the value
  - We identified and delivered the first 7 showcases to test the value of data in audit

- **Scale fast**
  - Build the operating model
  - We implemented a hub & spoke model and trained 300 auditors to become data driven

- **Harvest**
  - Data Driven Assurance
  - 35 cases per yr
  - 57 cases per yr

- **2017** 7 cases
- **2018** 13 cases
- **2019** 35 cases per yr
- **2020** 57 cases per yr
No change without challenges

LEARNINGS:
• Tone at the top, strategic change
• Auditor analysts drive scale up
• Carrot and the stick
• Integrated in Audit Procedures
• Benefit realization
• Data approval process

CHALLENGES:
• Conservatives
• Top-down hypotheses
• Next maturity Lab environment
• More positive assurance
USE CASE

Anomaly Detection: Unknown Patterns in Anti-Money Laundering
What is AML?

Source: UNODC
Known versus unknown patterns

Hypothesis A
Hypothesis B
Hypothesis C

Set of defined business rules

Johari window (1955)
Known versus **unknown** patterns

**Unknown Patterns in auditing Anti-Money Laundering**

What are the clients that show different behavior than others?  
What are clusters of clients that show similar behavior?  

Johari window (1955)
Unknown Patterns in auditing Anti-Money Laundering

**KNIME for non-techies**

---

**DATA ACCESS**
Select an input file. Prepare your data before using this workflow. Data preparation, e.g., cleaning the data, is not part of this workflow.

**INSPECT AND EXCLUDE COLUMNS**
Inspect the nominal and categorical columns. Choose what columns to exclude in further analyses.
- Columns with many missing values are likely to be less valuable for isolation forest.
- Columns with many categories are deemed less valuable for isolation forest. Try to group categories together, e.g., group countries in logical categories.
- Excluding columns from large datasets takes long. Alternatively, filter by using a columns filter before applying this component, or use smaller dataset as input.

**ANOMALY DETECTION**
Runs the anomaly detector. Each case is assigned an anomaly score between 0 and 1.
- The H2O anomaly detector is applied: this is an isolation forest method that does not need Python to be installed.
- Optional: Adapt the number of models and/or the number of levels (Tree depth) that the anomaly detector uses.
- The number of models parameter allows you to configure the number of decision trees composing the ensemble of trees (hence, the forest size). This must be a number up to 256. Higher numbers tend to give better results although they take longer to process. Default: the number of models is 100.
- The number of levels parameter allows you to configure the number of levels of a tree (hence, the tree depth). By default, the number of levels is 10.

**VISUALIZATION**
Calculation of the feature importance for the anomaly scores of each case/row.
To inspect all relevant features of an anomalous row, select this row, click ‘close and apply’ and run ‘visualize single entry’.

**OUTPUT**
Download the results including the anomaly score per case and feature importance.
- Do you want to inspect a specific anomaly (hence, row)? You could use a rule-based row filter and select a row with row number/row of original file. Example statement: $$\text{ROW}$$ = "Row778" => TRUE.

---

**Microsoft SQL Server Connector**
**DB Reader**

**Joiner**
**Excel Writer (XLS)**
Unknown Patterns in auditing Anti-Money Laundering

Under-the-hood...
The art of anomaly detection

Points to take into account

• Start small (<n variables)

• Understand the detected anomalies

• Garbage in, garbage out! Quality of the input...

• Very very important which features you import in the model

• You can have groups with the same anomaly scores → identify clusters

• Spot data quality issues
Thanks for listening, be involved!