Will it Blend?

Blending Enterprise-BI, Data Science, and Modern IT-Architecture – Lessons Learned

Dr. Matthias Nolte, March 7, 2018
Blending Data, Skills, and Technology
Blending ingredients to make the final product

Raw Data

(Data Lake, Staging, etc.)
Blending ingredients to make the final product

Data Collection

(Data Warehouse)
Blending ingredients to make the final product

Corporate DataStore

row-oriented
Blending ingredients to make the final product

Analytical DataStore
Blending ingredients to make the final product

Analytical DataStore

In Memory
Blending ingredients to make the final product

Data Frames for Machine Learning
Blending ingredients to make the final product

Dataset (usecase-specific)

(internal Data)
Blending ingredients to make the final product

Dataset (usecase-specific)

(external provider)
Blending ingredients to make the final product

(美味的)

Analytics
Variety of Analytical Tools to choose from
Main Concern: Is the End Product tasty?
Blending Skills
Assignment of tasks within the team (2008)

- Data-Modelling and Platform Management
- Report-design
- Dashboard-Development
- Ad hoc - Analysis
- Domain Knowledge - Customer Contact
- Machine Learning
Assignment of tasks within the team (2018)
Assignment of tasks within the team (2018)
Standardised Communication and Collaboration Framework
Our Approach: Disciplined Agile Delivery (DAD)

- Identify, prioritize, and select projects
- Initial Vision & Funding
  - Initial modeling, planning, and organization
  - Business Roadmap, Technology Roadmap
- Inception
  - One or more short iterations
  - Stakeholder vision
  - Proven architecture
- Construction
  - Many short iterations producing a potentially consumable solution each iteration
  - "viable product"
- Transition
  - One or more short iterations
  - Sufficient functionality
  - Production ready
  - Delighted stakeholders

Daily Coordination Meeting

Iteration

Initial Architectural Vision

Iteration planning session to select work items and identify work tasks for current iteration

Work Items

Highest-Priority Work Items

Consumable Solution

Iteration review & retrospective: Demo to Stakeholders, determine strategy for next iteration, and learn from experience

Funding & Feedback

Enhancement Requests and Defect Handling

Release solution into production

Operate and support solution in production

Initial Requirements and Release Plan

Daily

Initial model, planning, and organization

Work Items

Iteration Backlog

Tasks

Stakeholder input
IT Architecture
IT Architecture

Building Blocks

- the fruit analogy
Business Intelligence

PLATIN
(Plattform für Informationsanalyse)
**Business Intelligence**

- **Distribution/ Scheduling**
- **Report Development**
  - **BI Launch Pad**
  - **Ad-hoc Analysis**
  - **Detailed Analysis**
- **Data Services**
  - **Validate, Cleanse, Harmonize,…**
  - **General framework for ETL**
- **Data Profiling**

Additional services may include:
- **Load**
- **Transform**
- **Extract**
- **Information Steward**

**SAP**

**Dr. M. Nolte**
Technology Blending

BI and Advanced Analytics are fully integrated
Blending

SAP HANA and KNIME
Blending POC:

SAP HANA and KNIME

Use Case: – Prediction of loan default

1) Training Data und Population Data reside in Analytical Database (SAP HANA) (fictional data)
2) KNIME retrieves data from HANA
3) Training of two Algorithms
4) Automated best-model-selection
5) Generation of SQL
6) Direct Application of best model to the population in SAP-HANA
Data Prep

Training Data (SAP HANA)

Model Training (Decision Tree & Logistic Regression)

Database Connection
Table Reader

SAP HANA Connector

Missing Value
Partitioning

SAP HANA to KNIME
impute missing values
65% training
35% test

Berliner Sparkasse
Training Data Set

- **Training Data (SAP HANA)**
  - SAP HANA Connector
  - Database Connection Table Reader
  - Missing Value: impute missing
  - Partitioning: 65% training, 35% test
  - SAP HANA to KNIME

- **Model Training (Decision Tree & Logistic Regression)**

Table "database" - Rows: 5960

<table>
<thead>
<tr>
<th>Row ID</th>
<th>$s$ target</th>
<th>$s$ income</th>
<th>$s$ mortgage</th>
<th>$s$ value</th>
<th>$s$ reason</th>
<th>$s$ job</th>
<th>$D$ yoj</th>
<th>$D$ deroga</th>
<th>$D$ delinqu</th>
<th>$D$ dage</th>
<th>$D$ inquiries</th>
<th>$s$ chio</th>
<th>$D$ debitnc</th>
</tr>
</thead>
<tbody>
<tr>
<td>Row0</td>
<td>1</td>
<td>25860</td>
<td>39025</td>
<td>HomeImp</td>
<td>Other</td>
<td>10.5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>94.367</td>
<td>1</td>
<td>9</td>
<td>33.78</td>
</tr>
<tr>
<td>Row1</td>
<td>1</td>
<td>70053</td>
<td>68400</td>
<td>HomeImp</td>
<td>Other</td>
<td>7</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>121.833</td>
<td>0</td>
<td>14</td>
<td>33.78</td>
</tr>
<tr>
<td>Row2</td>
<td>1</td>
<td>15300</td>
<td>16700</td>
<td>HomeImp</td>
<td>Other</td>
<td>8</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>149.467</td>
<td>1</td>
<td>10</td>
<td>33.78</td>
</tr>
<tr>
<td>Row3</td>
<td>1</td>
<td>73750</td>
<td>101775</td>
<td>DebtCon</td>
<td>Other</td>
<td>8.922</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>179.766</td>
<td>1</td>
<td>21</td>
<td>33.78</td>
</tr>
<tr>
<td>Row4</td>
<td>0</td>
<td>97800</td>
<td>112000</td>
<td>HomeImp</td>
<td>Office</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>93.333</td>
<td>0</td>
<td>14</td>
<td>33.78</td>
</tr>
<tr>
<td>Row5</td>
<td>1</td>
<td>30548</td>
<td>40320</td>
<td>HomeImp</td>
<td>Other</td>
<td>9</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>101.467</td>
<td>1</td>
<td>8</td>
<td>37.114</td>
</tr>
<tr>
<td>Row6</td>
<td>1</td>
<td>1800</td>
<td>48649</td>
<td>HomeImp</td>
<td>Other</td>
<td>5</td>
<td>3</td>
<td>2</td>
<td>77.1</td>
<td>1</td>
<td>17</td>
<td>17</td>
<td>33.78</td>
</tr>
<tr>
<td>Row7</td>
<td>1</td>
<td>1800</td>
<td>28502</td>
<td>HomeImp</td>
<td>Other</td>
<td>11</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>88.766</td>
<td>0</td>
<td>8</td>
<td>36.885</td>
</tr>
<tr>
<td>Row8</td>
<td>1</td>
<td>2000</td>
<td>32700</td>
<td>HomeImp</td>
<td>Other</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>216.933</td>
<td>0</td>
<td>13</td>
<td>33.78</td>
</tr>
<tr>
<td>Row9</td>
<td>1</td>
<td>2000</td>
<td>73750</td>
<td>HomeImp</td>
<td>Sales</td>
<td>15</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>115.8</td>
<td>0</td>
<td>13</td>
<td>33.78</td>
</tr>
</tbody>
</table>

**Berliner Sparkasse**

Dr. M. Nolte 32
Model Training

Decision tree

Partitioning

65% training, 35% test

PMML To Cell

Decision Tree Learner

extract PMML

Decision Tree Predictor

train model

test model

Cross Joiner

JavaScript ROC Curve

merge

ROC AUC

Logistic Regression

PMML To Cell

Logistic Regression Learner

extract PMML

Regression Predictor

train model

test model

Cross Joiner

JavaScript ROC Curve

merge

ROC AUC
Generation of ROC AUC for Model Selection
Best-model Selection and generation of SQL
Applying the Model
**HANA-table with Targets**

Here is a table with fictional data:

<table>
<thead>
<tr>
<th>Row ID</th>
<th>Person</th>
<th>D P1</th>
<th>D P0</th>
<th>S Target</th>
</tr>
</thead>
<tbody>
<tr>
<td>Row 1</td>
<td>46539745</td>
<td>0.218</td>
<td>0.782</td>
<td>0</td>
</tr>
<tr>
<td>Row 2</td>
<td>18684763</td>
<td>0.465</td>
<td>0.822</td>
<td>0</td>
</tr>
<tr>
<td>Row 3</td>
<td>69951634</td>
<td>0.178</td>
<td>0.87</td>
<td>0</td>
</tr>
<tr>
<td>Row 4</td>
<td>3375942</td>
<td>0.13</td>
<td>0.87</td>
<td>0</td>
</tr>
<tr>
<td>Row 5</td>
<td>41877127</td>
<td>0.183</td>
<td>0.817</td>
<td>0</td>
</tr>
<tr>
<td>Row 6</td>
<td>56184929</td>
<td>0.242</td>
<td>0.758</td>
<td>0</td>
</tr>
<tr>
<td>Row 7</td>
<td>28291372</td>
<td>0.888</td>
<td>0.113</td>
<td>0</td>
</tr>
<tr>
<td>Row 8</td>
<td>71310714</td>
<td>0.215</td>
<td>0.785</td>
<td>0</td>
</tr>
<tr>
<td>Row 9</td>
<td>47971844</td>
<td>0.415</td>
<td>0.663</td>
<td>0</td>
</tr>
<tr>
<td>Row 10</td>
<td>36784339</td>
<td>0.138</td>
<td>0.763</td>
<td>0</td>
</tr>
<tr>
<td>Row 11</td>
<td>79606932</td>
<td>0.132</td>
<td>0.858</td>
<td>0</td>
</tr>
<tr>
<td>Row 12</td>
<td>55826239</td>
<td>0.34</td>
<td>0.663</td>
<td>0</td>
</tr>
<tr>
<td>Row 13</td>
<td>70662294</td>
<td>0.34</td>
<td>0.763</td>
<td>0</td>
</tr>
</tbody>
</table>

**Inject SQL to**

- Import population data from HANA to KNIME.
YES, IT BLENDS!
Next step
Development of Custom HANA Nodes

Berliner Sparkasse

KNIME

SAP

SAP HANA Nodes

COMING SOON

OPEN SOURCE

Dr. M. Nolte
Let’s stay in touch

@d4t4v1z
Dr. Matthias Nolte
(views are my own)