The Data Innovation Project NOGAuto

Lorenz Helbling, Cindia Duc Sfez
Group of Experts on Business Registers
September 28, 2022
Agenda

1. NOGAuto context and description
2. Today vs future
3. Process
4. Measures of success
5. Workflow
6. Interface
Leading consequently to an overall quality improvement of all enterprise statistics.

01 New FSO data innovation strategy

02 Pilot project

03 Support the collaborators in their daily work

04 Make the classification of NOGA codes more accurate
Description of the NOGAuto project

The creation of a program that can automatically predict companies' NOGA codes using “Machine Learning” (ML) techniques with a coding quality that is equal or higher than the manual coding currently performed.

Reduction of the interpretation factor

Standardisation of codification

Time saving and optimisation

Improvement
Today vs. Futur

**Input:** Activity description of a company

**Coding**

- Quality control by sampling (~ 30%)

**Processing of requests**

---

**Futur**

**Input:** Activity description of a company

**Coding**

- Quality control by sampling (~ 99.9%)

**Processing of requests**

---

- September 2022
Data innovation as a process

Phase 1
- Use Case Backlog
- Prototype (CRISP-DM)
  - Data Preparation
  - Data Understanding
  - Business Understanding
  - Validation
  - Modeling

Phase 2
- Minimal Viable Product (MVP)
  - Design Adjustment
  - Architecture Adjustment
  - Realising
  - Continuous Testing

Phase 3
- Production
  - Release
  - Operating Activity
  - Monitoring
Measures of success

<table>
<thead>
<tr>
<th></th>
<th>Phase 1</th>
<th>Phase 2</th>
<th>Phase 3</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Prediction ML</strong></td>
<td>75 %</td>
<td>85 %</td>
<td>95 %</td>
</tr>
<tr>
<td>system-centered</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Predictive accuracy</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>(1 – generalisation</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>error)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Documentation</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Monitoring</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td><strong>Domain user-</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>centered</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Interface</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Performance</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Daily use</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Added value for the</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>user</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Software used

renku

RStudio
Workflow of the «NOGAuto ML system»
1. Data Collection

- Internal Database
  - Business Enterprise Register BER (commercial registers, VAT, customer transactions (borders))
  - Over 1'000'000 observations with 56 variables each

2. Data Preparation

- Language detection (e.g. naive Bayes, deep learning, N-gram, root cause analysis)
- Text mining & natural language processing (NLP) (e.g. one-hot encoding, Text2vec, Word2vec, t-SNE, word clouds)

3. Modeling

- GBM: Gradient Boosting Machine
- Three different models trained according to the language
- Cross-validation
1. Data Collection

- Internal Database
- Business Enterprise Register BER (commercial registers, VAT, customer transactions (borders))
- Over 1,000,000 observations with 56 variables each

2. Data Preparation

- Language detection (e.g. naive Bayes, deep learning, N-gram, root cause analysis)
- Text mining & natural language processing (NLP) (e.g. one-hot encoding, Text2vec, Word2vec, t-SNE, word clouds)

3. Modeling

- GBM: Gradient Boosting Machine
- Three different models trained according to the language
- Cross-validation
1. Data Collection

- Internal Database
- Business Enterprise Register BER (commercial registers, VAT, customer transactions (borders))
- Over 1'000'000 observations with 56 variables each

2. Data Preparation

- Language detection (e.g. naive Bayes, deep learning, N-gram, root cause analysis)
- Text mining & natural language processing (NLP) (e.g. one-hot encoding, Text2vec, Word2vec, t-SNE, word clouds)

3. Modeling

- GBM: Gradient Boosting Machine
- Three different models trained according to the language
- Cross-validation
NOGAuto Modeling – 2 Digits predictions

2 Digits Prediction

<table>
<thead>
<tr>
<th>Real NOGA_2D</th>
<th>1st Prediction</th>
<th>1st %</th>
<th>2nd Prediction</th>
<th>2nd %</th>
<th>3rd Prediction</th>
<th>3rd %</th>
<th>Match (=1) vs No Match (=0)</th>
</tr>
</thead>
<tbody>
<tr>
<td>01</td>
<td>01</td>
<td>99.39%</td>
<td>03</td>
<td>0.28%</td>
<td>02</td>
<td>0.11%</td>
<td>1</td>
</tr>
<tr>
<td>47</td>
<td>47</td>
<td>86.32%</td>
<td>10</td>
<td>8.71%</td>
<td>46</td>
<td>1.96%</td>
<td>1</td>
</tr>
<tr>
<td>96</td>
<td>96</td>
<td>99.87%</td>
<td>47</td>
<td>0.06%</td>
<td>56</td>
<td>0.014%</td>
<td>1</td>
</tr>
<tr>
<td>01</td>
<td>86</td>
<td>78.79%</td>
<td>01</td>
<td>19.60%</td>
<td>47</td>
<td>0.19%</td>
<td>1</td>
</tr>
<tr>
<td>86</td>
<td>86</td>
<td>41.00%</td>
<td>74</td>
<td>14.32%</td>
<td>70</td>
<td>13.70%</td>
<td>1</td>
</tr>
<tr>
<td>47</td>
<td>47</td>
<td>93.15%</td>
<td>70</td>
<td>3.71%</td>
<td>96</td>
<td>1.04%</td>
<td>1</td>
</tr>
<tr>
<td>49</td>
<td>96</td>
<td>16.37%</td>
<td>49</td>
<td>15.68%</td>
<td>69</td>
<td>7.51%</td>
<td>1</td>
</tr>
<tr>
<td>01</td>
<td>01</td>
<td>99.89%</td>
<td>03</td>
<td>0.03%</td>
<td>96</td>
<td>0.007%</td>
<td>1</td>
</tr>
<tr>
<td>01</td>
<td>01</td>
<td>99.89%</td>
<td>03</td>
<td>0.002%</td>
<td>02</td>
<td>0.001%</td>
<td>1</td>
</tr>
<tr>
<td>01</td>
<td>01</td>
<td>99.89%</td>
<td>03</td>
<td>0.03%</td>
<td>86</td>
<td>0.009%</td>
<td>1</td>
</tr>
<tr>
<td>01</td>
<td>01</td>
<td>99.36%</td>
<td>03</td>
<td>0.515%</td>
<td>96</td>
<td>0.013%</td>
<td>1</td>
</tr>
<tr>
<td>49</td>
<td>49</td>
<td>98.74%</td>
<td>88</td>
<td>0.38%</td>
<td>85</td>
<td>0.13%</td>
<td>1</td>
</tr>
<tr>
<td>85</td>
<td>85</td>
<td>85.35%</td>
<td>90</td>
<td>10.38%</td>
<td>47</td>
<td>1.044%</td>
<td>1</td>
</tr>
<tr>
<td>01</td>
<td>01</td>
<td>99.89%</td>
<td>03</td>
<td>0.03%</td>
<td>96</td>
<td>0.007%</td>
<td>1</td>
</tr>
</tbody>
</table>

Dataset Predictive accuracy (top 3)

- French ~ 99.12%
- German ~ 96.16%
- Italian ~ 98.25%
# NOGAuto Modeling – 4 Digits predictions

## 4 Digits Prediction

<table>
<thead>
<tr>
<th>Real NOGA Code</th>
<th>1st Prediction</th>
<th>In %</th>
<th>2nd Prediction</th>
<th>In %</th>
<th>3rd Prediction</th>
<th>In %</th>
</tr>
</thead>
<tbody>
<tr>
<td>8412</td>
<td>8891</td>
<td>53.85%</td>
<td>8899</td>
<td>7.112%</td>
<td>8810</td>
<td>4.48%</td>
</tr>
<tr>
<td>0143</td>
<td>0143</td>
<td>13.81%</td>
<td>0150</td>
<td>13.62%</td>
<td>0121</td>
<td>8.602%</td>
</tr>
<tr>
<td>8621</td>
<td>8621</td>
<td>24.05%</td>
<td>8622</td>
<td>19.03%</td>
<td>8690</td>
<td>4.455%</td>
</tr>
<tr>
<td>0141</td>
<td>0141</td>
<td>10.54%</td>
<td>0150</td>
<td>10.42%</td>
<td>0121</td>
<td>7.594%</td>
</tr>
<tr>
<td>8690</td>
<td>8690</td>
<td>52.246%</td>
<td>9602</td>
<td>5.707%</td>
<td>9609</td>
<td>5.373%</td>
</tr>
<tr>
<td>8621</td>
<td>8621</td>
<td>51.30%</td>
<td>8622</td>
<td>4.221%</td>
<td>8559</td>
<td>3.22%</td>
</tr>
<tr>
<td>0149</td>
<td>0149</td>
<td>31.91%</td>
<td>0150</td>
<td>14.55%</td>
<td>0143</td>
<td>1.639%</td>
</tr>
<tr>
<td>2341</td>
<td>2341</td>
<td>24.078%</td>
<td>8552</td>
<td>19.55%</td>
<td>8559</td>
<td>6.423%</td>
</tr>
<tr>
<td>0150</td>
<td>0141</td>
<td>19.24%</td>
<td>0111</td>
<td>7.834%</td>
<td>0150</td>
<td>6.713%</td>
</tr>
<tr>
<td>8622</td>
<td>8621</td>
<td>58.545%</td>
<td>8622</td>
<td>8.796%</td>
<td>9900</td>
<td>8.482%</td>
</tr>
</tbody>
</table>

## Dataset Predictive accuracy (top 3)

<table>
<thead>
<tr>
<th>Language</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>French</td>
<td>~ 88%</td>
</tr>
<tr>
<td>German</td>
<td>~ 40%</td>
</tr>
<tr>
<td>Italian</td>
<td>~ 95%</td>
</tr>
</tbody>
</table>
Example of activity descriptions

1. **FRENCH**
   La société a pour but les conseils et formations en matière de finance, marketing et développement commercial; elle peut acquérir et administrer des participations dans d'autres entreprises ainsi qu'exercer des activités de support du management (pour but complet cf. statuts).

2. **GERMAN**
   Ausführung aller Arbeiten im Fachgebiet Schreinerarbeiten, sowie den Handel mit Produkten der Holzindustrie.

3. **ITALIAN**
   La produzione e la vendita di vini e distillati, aceto e olio d'oliva sia in Svizzera che all'estero così come ogni altra attività atta a conseguire lo scopo sociale. La società potrà inoltre acquistare, vendere e amministrare beni immobili.
UI: User Interface
Next steps

1. Web-application deployment
2. Re-train models
3. Send to production
Other potential deployment

NACE Revision (2024-2025)

• Potential supporting tool in the double codification for the collaborators for the complex cases namely companies with 1:n relations.

• The system will use only the observations of the dataset which are in the n categories and will train a ML model for the company being recoded.

Application on other nomenclatures

• Use the same methodology on different nomenclatures such as the profession ISCO nomenclature.

• First results are promising
Thank you for your attention

Questions?