UNECE HLG-MOS ML Project Classification & Coding Theme Report

Claus Sthamer Claus.Sthamer@ONS.gov.uk
The C&C Theme

• C&C, along with Edit & Imputation and Imagery themes, is part Work Package 1

• C&C objective: This theme was selected from the GSBPM as one of the processes suitable for Machine Learning
In the terminology of machine learning, classification is considered an instance of supervised learning, i.e., learning where a training set of correctly identified observations is available.”

Challenge of manual C&C?

- Very labour intensive
- Repetitive
- Lengthy process – delay in statistical output
- Inconsistent – level of experience of coders might differ
- Deterministic/rule-based/word matching systems are difficult to build and maintain, rules/reference text entries have to be updated frequently
## Pilot Studies – How can Machine Learning help?

<table>
<thead>
<tr>
<th></th>
<th><strong>Organization</strong></th>
<th><strong>Survey/Activity</strong></th>
<th>Primary Field</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>BLS – USA</td>
<td>Survey of Occupational Injuries and Illnesses</td>
<td>Workplace Injury – SOC, OIICS, 6 codes</td>
</tr>
<tr>
<td>2.</td>
<td>Stats Canada</td>
<td>Canadian Community Household Survey</td>
<td>Occupation &amp; Industry – NAICS, NOC</td>
</tr>
<tr>
<td>4.</td>
<td>INEGI – Mexico</td>
<td>Household Income and Expenditure</td>
<td>Occupation &amp; Economic activity - SCIAN, SINCO</td>
</tr>
<tr>
<td>5.</td>
<td>Statistiek Vlaaderen – Belgium</td>
<td>Sentiment of Twitter Data</td>
<td>Positive/Negative</td>
</tr>
<tr>
<td>7.</td>
<td>Statistics Poland</td>
<td>Web scraped food products</td>
<td>Food description - ECOICOP</td>
</tr>
<tr>
<td>8.</td>
<td>IMF</td>
<td>Catalogue of Time Series - CTS</td>
<td></td>
</tr>
</tbody>
</table>
Pilot Studies Objectives

Quality – Efficiency – Timeliness - Accuracy
Pilot Studies - Insights

- Data Requirements: Golden Data Set – Ground Truth
- Algorithms:
  - SVM & XGBoost, Random Forest, FastText performed well
  - Neural Network best result
- IT hardware:
  - Normal desktop/laptop used by 5 pilot studies
  - Cloud computing used by one
  - 4 x Graphical Processing Units with 3584 cores used to run a neural network
- Quality Measures used
  - Accuracy
  - Recall
  - F1-score
Value added by ML for C&C

• Auto-coding can be achieved, but not for 100%

• ML/Human work together to get best results
  – Prediction threshold – auto-coding where ML is ‘sufficiently’ confident
  – Human coding of minority classes and low confident predictions, >95% Accuracy

• Faster processing than manual

• Increased data consistency
Challenges/Blockers

• 3 of 8 are in production

• The other 5:
  – High demand for analysts
  – Volume and quality of training data
  – Accuracy
  – Not enough resource to progress the pilot study
  – Cost of IT systems - prevent usage of Neural Networks
Thank you